ADAPTIVE LEARNING REDEMPTION RATE PREDICTION MODEL

by

Camila Helena Ölund Matos

Plan for Dissertation of the Master in Modeling, Data Analysis and Decision Support Systems

Academic Advisors

Professor Doutor João Gama

Professor Doutor Jorge Valente

Internship Advisor:

Liliana Bernardino

Faculdade de Economia

Universidade do Porto

2017/18

Acknowledgements

The development of this dissertation was a result of collaborative efforts.

I would like to express my appreciation to Professors Dr. João Gama and Dr. Jorge Valente for their advice, support and encouragement.

Furthermore, I would like to thank the Continente Card's Management Board for conceding the opportunity to develop my dissertation for a problem with real-world settings. I owe a special thanks to Liliana Bernardino, Ana Freitas, Patrícia Castro and Sílvia Cunha for their cooperation and assistance with this research.

A heartfelt appreciation goes out to my family and friends which inspired me to try harder and do better.

Aos meus pais, Carlos e Ingela.

Abstract

Nowadays, the implementation of Direct Marketing strategies has attained an increased relevance over Mass Marketing campaigns, for allowing the implementation of more successful and profitable marketing strategies, by customizing offers for segmented customer groups.

The groundwork of this dissertation is set upon the retail market, in which is the creation of loyalty programs, such as the use of customer cards, allows the collection and treatment of demographic, customer behavior and transactional data, which favors the implementation of Direct Marketing strategies. Using this information, customized and segmented campaigns are created for different customer groups based on the business strategies and directives.

In this dissertation the focus is set upon promotional coupon direct marketing campaigns, sent to customers as stimulus. The implementation of this sort of campaign urges the need to predict the percentage of customer adhesion to these campaigns, since this knowledge provides strategic decision support regarding stock management, strategic outline and sales forecast.

However, predicting the percentage of redeemed promotional coupons (campaign adhesion) is not the sole ambition of this project. It is intended to create an Adaptive Machine Learning model framework, so time-dependent changes occurring in the input variables of the model are monitored and the model is responsive to them, in real time.

To achieve it, an in-depth study of the state of art was carried out, as well as the study of the business' specificities. Then, the implementation of the Data Mining and Adaptive Machine Learning methods were carried out regarding the Business' and Data Mining's defined goals and success criteria.

Keywords: Adaptive Machine Learning, Data Mining, Direct Marketing, Coupon Redemption Rate Prediction.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Framework

The fast evolution the retail industry and consumer behavior have undergone in the last few years has brought some new challenges to retailers operating in competitive markets.

Alongside the increase in competition came a more defiant customer- showcasing a different kind of behavior than before- making more conscious and informed choices and being more demanding regarding the quality of the products and services available, thus becoming less loyal than one used to be (Verhoef, 2002).

Therefore, the need to define new strategies and to set ambitious goals has grown, making retail companies more aware of the importance of analytics as an opportunity to add real value to both the business and the consumer, by providing a better understanding of them (Kolter, 2005). The resulting decision support systems, backed up by predictive and prescriptive analytics, lead towards more sustained, assertive and effective decisions in the future and allow companies to trigger situations which meet its goals, assuring better results (Davenport,2010). This way, retailers benefit more from the development of methodologies and models that forecast future events, rather than simply analyze what and why did something happen in the past (Olson, 2009).

In parallel with the evolution of the retail and data analysis markets, a new view over the customer came to light. Top management started to recognize the importance of customer satisfaction and loyalty and perceive it as the main advantage over the competitors (Bateson, 1997). Thus, in order to build long-term satisfactory relationships with customers and acquire a lasting competitive advantage, companies have to establish a dialog by employing marketing strategies that provide value for the customers (Kolter, 2005).

Nowadays, marketing and promotions carried out in the business market are based, essentially, in two approaches: Mass Marketing and Direct Marketing.

While Mass Marketing uses mass media and mass channels to disseminate a message equally and uniformly to all its customers and potential future customers; direct marketing favors the customization of interactions for specific groups of customers classified according to their characteristics and preferences (Kotler,1989).

The implementation of direct marketing has grown over the last two decades, triggering both corporate and academic research interest, which might be explained by the fact that it can aid businesses and their departments to conduct their campaigns, management and results more efficiently and effectively, since it is associated to higher returns on investment (Eisenstein, 2002). Consequently, direct marketing campaigns are increasingly sought to optimize the distribution of different messages to different customer segments, which are created from the records of consumer behavior (Davenport, 2010).

For this purpose, loyalty programs (such as loyalty cards) have emerged as a rewards program offered by a company to customers who frequently make purchases, and which allow the collection of the customer's socio-demographic, behavioral and transactional data. This customer activity data will update the marketing teams' knowledge on customer behavior and influence future decisions, by enabling the creation of more suitable and valuable offers to the consumer and providing better products and services (Davenport, 2010).

Since it is considered an important promotional tool for companies, the need to improve the efficiency and articulate the result of these promotional campaigns in store with several business areas is a constant concern for companies. Hence, the ability to predict the customers' adhesion to these offers is, therefore, as much a goal as it is a necessity (Baesens, 2002).

To achieve that goal, the customer's response to a stimulus can be modelled and used to predict the probability and percentage of adhesion to a certain promotional offer, by the target group of that same offer.

Summarizing, predictive models in direct marketing use the characteristics of the clients, based on the records of their past behavior and campaign participation, in an effort to aid the decision-making process of several business departments (Davenport, 2010), regarding the outcome of the promotional campaigns.

1.2 Problem Discussion and Definition

Since mass marketing has become less and less efficient at acquiring and retaining customers, direct marketing campaigns have become a more efficient strategy (Kotler, 1989). This type of strategy provides companies with the ability to target customers with whom a direct interaction will be established and enables a more accurate forecast of the campaign's reach and outcome. Therefore, direct marketing is carried out by mailing specialized offers - such as product offers, ticket promotions and general communication - directly to a defined public.

The direct marketing strategy under analysis on this thesis has promotional mailing coupon campaigns as focus, more specifically for the largest Continente Card partner, Continente. These coupons can be redeemed in both physical and online retail stores for discounts and will work as stimulus, sent directly to the customers by different channels (e.g. letter, newsletter, mobile app).

The retail company subject to this dissertation is Sonae MC, which takes part in a 10-year-old loyalty program of an ecosystem of brands with 3,5 million active accounts. This loyalty program is materialized in the form of a card which enables the tracking of transactions and provides the access to personalized promotions, events, specific partnerships and a permanent discount in all transactions. The resulting database enables the implementation of direct marketing strategies used to leverage the business of its partners (Davenport, 2010). This data will also be used as an input to ultimately build the Adaptive Redemption Rate Prediction Model of Promotional Coupons.

The relevance of applying technical and analytical efforts to predict the Redemption Rate of Promotional Coupons is that this analysis can provide valuable information for the business, in instances such as: to make more adequate management of the stocks by avoiding stock outs and spoilage, to make strategic adjustments to the segmented groups if necessary, to adjust sales forecasts (Granja, 2017) and to optimize the distribution of mailing coupons.

However, predicting the Redemption Rate of Promotional Coupons is not the sole ambition of this project. Noting that the retail market is a highly competitive one (Bateson, 1997)., which requires increasing marketing costs, a high interaction with

customers through direct marketing strategies and a quick and constant adaptability to market changes and shifts, it is important to predict the reach and outcome of these promotional campaigns to aid the decision-making process (Davenport, 2010).

In fact, Machine Learning systems are typically open loop, meaning there is no direct link between the Machine Learning system and the collection of data (Gepperth & Hammer, 2016). On the other hand, Adaptive Machine Learning not only selects experiments to discriminate between contending hypotheses but also carries out these experiments in the learning domain, changing its behavior at the time it is performed. (Bryant, 2000).

As such, the goal of this dissertation is also to create an adaptive model framework, so time-dependent changes are considered, enabling the predictive model to adjust without human interaction (Rico-Juan & Iñesta 2014). The importance of this adaptation has to do with the fact that data entities vary- for instance, over time- and monitoring them will improve the results and may even be necessary on an on-going basis if the changes are significant (Bryant, 2000), lest the learned system becomes "aged" and not responsive to new changes.

In other words, it is intended to create a predictive model with adaptive configurations capable of learning, adapting and making improvements in the ongoing process i. e. to build a model capable of monitoring the performance of events by automatically deeming what is the norm and predicting when failures are likely to occur, so that when trends change or new datasets are introduced, there is no need for human interaction. Hence, from the steady-state, the Machine Learning algorithm will learn and deem what is the norm, and when deviations occur, Adaptive Machine Learning will flag this change by monitoring the model´s parameters, and quickly learn the new model.

Synthesizing, this thesis has two main goals: not only to predict the probability of redemption of a certain promotional coupon from those sent to the target group of customers, but also to implement an adaptive learning optimization algorithm, so the model is able to perform adjustments to shifts in the market, in real time.

1.3 Research Questions

The scope of this dissertation is build an adaptive learning model to identify the probability of obtaining a positive response to a promotional offer by a target group of customers; using the historical records of transactions of the clients registered in the Continente Card database as input. Throughout this process, the following issues will be addressed:

What are the chances of stimulated customers making use of a promotional offer in the next marketing campaign, based on their transactional data and interactions in previous marketing campaigns?

What percentage of promotional coupons sent per letter, for target group, that will effectively be redeemed?

Does the implementation of an adaptive learning model bring improvements to the ongoing processes, by improving the accuracy of the existing static model?

2 State of the Art

2.1 Data Mining & Adaptive Machine Learning

The exponential growth of accessible data and the consequent increase in complexity of data structures has prompted the use of more sophisticated technical and analytical methods by organizations (Olson, 2009), whichever the domain: Business and Marketing, Finance and Investment, Medical, Scientific Field, among others.

In fact, raw data by its own hardly ever delivers direct benefits. Instead, its true value lies in the extraction of useful information and knowledge. Thus, though being a challenging process, maximizing the gain of data by implementing analytical and optimization processes can be instrumental in competitive markets to add real value to an organization, by providing a better understanding of the data source, the business, the market and the consumer and enhancing the use of analytics to aid the process of strategic decision support (Davenport, 2010).

In this context, it is relevant to disclose the concept of Data Mining (hereinafter referred to as DM). DM is a process which attempts to draw conclusions and extract knowledge contained in huge amounts of data. In other words, it is "(...) a method directed to the discovery of hidden messages, such as trends, patterns and relationships in the data" (Hsu & Ho, 2012). Thus, DM consists of the articulation of several well-established areas of interest, such as traditional data analysis, artificial intelligence and automatic learning (Nisbet et al., 2009).

Data Mining methods infer structural information from given data. Most of the current applications restrict to the classical open-loop setting, where data is given prior to training, and training can rely on the assumption that the data and its underlying structure are static (Gepperth & Hammer, 2016).

Adaptive learning, in contrast, refers to the situation of continuous model adaptation based on a constantly arriving data stream and on a priori defined reward mechanism (Ruvolo & Eaton, 2013). Hence, the close loop nature of AML - where the solution tracks the changing environment - renders the solution automatically, removing the need for

human intervention (Hammer & Toussaint, 2015). Further, online learning becomes necessary in interactive scenarios where training examples are provided based on human feedback over time (Rico-Juan & Iñesta, 2014).

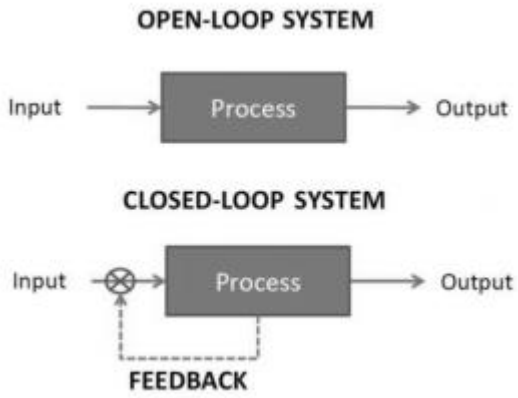

Figure 2.1 - Open-Loop system vs Closed Loop System

Therefore, for closed-loop cases, Adaptive Machine Learning will require online processing so that each data input is processed as it arrives (Kumar & Varaiya, 2015).

As such, the DM algorithm will learn the norm from the steady-state, then the Adaptive Machine Learning will flag the existent deviations and quickly learn and adapt the new model, while characterizing the change in terms of new model parameters i.e., it will not only detect that that a change as happened but also the nature of the change.

This way, it is relevant to describe how the data mining process is applied. According to (Jiawei et al., 2012), the data mining process involves the following steps: understand the domain of the application and formulate the problem, collect and pre-process the data, knowledge extraction and "hidden" patterns extraction, the interpretation and evaluation of the results and to apply the knowledge in a context of practical use.

While facing a Data Mining task, different data analysis techniques with different goals can be applied to achieve the desired results. These tasks can be classified into two categories: descriptive (Clustering, Association rule discovery, Sequential pattern discovery) and predictive (Classification, Regression, Deviation (outlier) detection). The

descriptive ones characterize the general properties of the data. Predictive tasks attempt to draw inferences from currently available data to make predictions (Jiawei et al., 2012).

2.2 Direct Marketing & Response Models

Information and knowledge gained through Data Mining processes can be instrumental for Marketing purposes (Verhoef, 2002)., which combined with the adaptive machine learning's ability to learn and adapt makes it the optimal choice for improvements in ongoing processes, marketing campaigns and continuous customer service improvements.

This is generally achieved through the collection, preparation and processing of huge amounts of data regarding customer behavior, individual or group interests, habits, preferences and demand (Davenport, 2010).

As mentioned, the scope of this project is focused in a type of Direct Marketing Campaign which aims, among other things, to reinforce customer engagement and loyalty to the brand by establishing and maintaining a direct relationship with the customers (Sérgio Moro et al., 2011).

Customers chosen for a specific offer and/or service are selected from the database considering different types of information such as demographics, personal data and purchase history (van Geloven, 2002). Thus, response models are created to identify which customers or prospects are more likely to respond to a direct stimulus and its main purpose is to increase revenue in future campaigns in relation to the investment made (Shin & Cho, 2006).

These models are called Response Models, where the dependent variable is the simple positive response to the request or not (Verhoef et al., 2002). Therefore, these models are usually formulated as a binary classification problem where customers are divided into two classes: those who respond to the request and those who do not.

The importance of such a method is that the information obtained by response models allows businesses to improve response rates by tracking and comparing the customer's direct marketing response results to a promotional action, and to entice the cutback of advertising costs by customizing offers to a more specific target group and casting aside ineffective channels (Baesens, 2002). It can also be used, for example, to decide which products and offers should be suggested to target groups, based on the likelihood or propensity of each customer to respond to an offer.

Response models have been proven as being an advantageous tool in fine-tuning direct marketing strategies since even small improvements obtained through modeling can lead to substantial financial gains (Elsner et al., 2004).

2.3 Review on Data Types

As in any modelling project, the selection of attributes that will be used as explanatory variables of the response model, is a crucial step. The quality of the field data used is the foundation of successful modelling (Vaughan, 2003).

Two types of data are generally used in direct marketing models: external and internal databases.

External databases contain geographical (such as home address), demographic (such as gender and age), lifestyle (such as buying habits) and socioeconomic conditions (such as wage), which are considered external data (Poel, 2003). This data can be retrieved from surveys or demographic studies such as the Census.

The second type of data- internal data- concerns customer behavior and interactions with marketers, which can be retrieved from transaction history, customer response to a stimulus, and web browsing logs. Internal databases provide the most relevant information regarding customer behavior (Setnes & Kaymak, 2001) and seldom current investigations focus exclusively on external data.

Moreover, data history can usually be translated into attributes based on the Recency, Frequency and Monetary (RFM) method (Poel, 2003), commonly used in retail to analyze

customer value and is mostly used for direct marketing modeling (Ramaswamy et al., 1998).

The RFM variables were firstly referred and identified by Cullinan & GJ (1977), which stand for three dimensions: Recency (measures the time that has passed since the last purchase), Frequency (measures the number of purchases made by the customer in a certain period) and Monetary (measures the monetary value associated with the purchases of a customer in a certain period).

Since then, several articles have used the RFM variables, reinforcing the importance and predictive power of these variables for response models to promotional offers, since the RFM information is used to estimate the likelihood of customers buying a certain product. Customer response or feedback to marketing campaigns is also a goal of this work and is used to discover and understand customer behavior (Viaene et al., 2001).

However, other factors are believed to influence the purchase decision of a product such as the characteristics of the product and of the promotional offer. Regarding the later, these factors spawn from its presentation, such as the presence of illustrations, the size of the envelope, paper type, or letter format (Ramaswamy et al., 1998), to the way the offer is carried out, i.e. concerning the channel used to establish contact: letter by mail, an email, text message via phone or a recommendation on a website.

As mentioned, the development of data collection and storage techniques has prompted the growth of available information in companies and it is common to have client records with hundreds of attributes. Yet, the indiscriminate use of attributes as inputs adds unnecessary complexity to the model and influences machine learning's ability to seek and find patterns in data, by bringing noise and leading to errors.

As a result, it is then necessary to select the most relevant explanatory variables of customer description and behavior. The feature selection methods and data preparation processes are elucidated in the Data Extraction and Description and Data Preparation chapters. The number of variables is thus reduced substantially.

2.4 Adaptive Learning Logistic Regression

The development of this dissertation followed the CRISP-DM structure. As such, the process was divided into six major steps: Business Understanding, Data Understanding, Data Preparation, Modelling, Validation & Evaluation and Deployment.

This chapter aims to present, in a first phase, some introductory concepts and, in another phase, the enumeration and description of the adaptive learning method proposed at this stage of the project.

Adaptive learning methods are a powerful tool that can perform optimization operations with minimal human intervention (Cui, Wong, & Lui, 2006).

The paradigm of the adaptive learning algorithms underlying this project is supervised learning which is directly related to prediction, while unsupervised learning is more related to the discovery of patterns in a data set (Stimpson & Cummings, 2014). In supervised learning, three sets of data are presented, the train, validation and test sets (Winandy, Borges Filho, & Bento, 2007).

The evaluation of a supervised algorithm is usually performed through the analysis of the predictor in the classification of new examples; that is, examples which were not used in the training set (Gama, 2015). The evaluation of the predictive models will allow a comparison of methodologies so the most adequate and high performing prediction techniques, in terms of measuring the overall prediction accuracy, are found.

2.4.1 Business Modelling Requirements

The methodology choice conciliated the state of the art research with Sonae's business requirements explored in the next chapter.

Even though several classification methods have been proposed in literature for Direct Marketing Response Modeling purposes such as Statistical Methods (Suh et al., 1999)

and, more recently, Machine Learning methods; including Decision Trees (Olson & Chae, 2012, Moro et al., 2014), Neural Networks (Moro et al., 2014, Olson & Chae, 2012, Potharst et al., 2001, Kim & Street, 2004, Suh et al., 1999, Yu & Cho, 2006), Bayesian Networks (Baesens et al., 2002, Wong, 2016), Genetic Algorithms (Kim & Street, 2004), Support Vector Machines (Javaheri et al., 2013, Shin & Cho, 2006, Yu & Cho, 2006), Naive Bayes (Ling & Li, 1998), and Meta-Heuristics (Coelho et al., 2017), in this dissertation the focus was set on Logistic Regression (Deichmann et al., 2002, Moro et al., 2014, Olson & Chae, 2012).

The reasoning to back this decision was the fact that one of the major goal for Sonae regarding this project was to optimize their existent Logistic Regression model. To facilitate the comparison and assess the model's improvement, it was decided to maintain a Logistic Regression Model.

This decision is also backed by many articles which refer to logistic regression as the most appropriate method due to its simplicity and interpretability (Hosmer & Lemesbow, 2000, Sen & Srivastava, 1990). In fact, other alternatives indicated by literature, such as Neural Networks, offer black-box solutions, which deviate from the company's goals since the model's interpretability is imperative.

On the other hand, the optimization algorithm selected to build the adaptive model was Stochastic Gradient Descent, widely popular in literature because of its inherent simplicity (Mathews, 1993).

Stochastic Gradient Descent one of the three variants of Gradient Descent: Batch Gradient Descent (BGD), Mini-Batch Gradient Descent (M-BGD) and Stochastic Gradient Descent (SGD).

A table of comparison of the three approaches was elaborated, revealing the upsides and downsides of each one of them:

	Upsides	Downsides
Stochastic Gradient Descent	Simple to understand and implement. An immediate insight into the performance of the model and the rate of improvement is given by the frequent updates. The update frequency may result in faster learning. The noisy update process may allow the model to avoid local minima.	Updating the model so frequently is computationally expensive. The frequent updates may result in a higher variance over training epochs i.e., a noisy gradient signal may induce the model error to jump around. The noisy learning process down the error gradient can also hinder the process of settling on an error minimum for the model.
Batch Gradient Descent	Fewer updates to the model are performed. The decreased update frequency results in a more stable error gradient and may result in a more stable convergence on some problems. Since the calculation of prediction errors and the model update are separate processes, the algorithm performs parallel processing-based implementations.	The more stable error gradient may result in premature convergence of the model to a less optimal set of parameters. The updates at the end of the training epoch require the additional complexity of accumulating prediction errors across all training examples. Its implementation requires the entire training dataset in memory and available to the algorithm. Model updates and training speed may become very slow for large datasets.
Mini-Batch	The higher update frequency than batch gradient descent allows a more robust convergence, avoiding local minima. The batched updates provide a computationally more efficient process than stochastic gradient descent. The batching allows both the efficiency of not having all training data in memory and algorithm implementations.	It requires the configuration of an additional hyperparameter for the learning algorithm: the size of the mini-batch. Additional complexity is added to the model since the updates at the end of the training epoch require the error information to be accumulated across mini-batches of training examples.

Table 2.1- Upsides and Downsides of the Gradient Descent Approaches based on (Bottou, 2012).

The reasoning backing this choice has to do with the amount of data used to compute the gradient of the objective function, which is operated differently for each one of the three approaches. In fact, a trade-off between the accuracy of the parameter update and the time it takes to perform an update is made depending on the amount of data at disposal (Bottou, (2012).

Since the database used in the development of project is massive and the business goals discussed in the next chapter are both to ensure an accurate model and to replicate the model for other business partners of the Continent Card in case of success, the major computational and memory expenses and time-effort consumption on the ongoing processes were considered in the decision-making process.

Therefore, SGD was chosen to the detriment of the other two methods because, since it performs one update at a time, it does not perform redundant computations for large datasets, which makes it much faster and also able to be used to learn online.

In the following sub-sections, Logistic Regression and Stochastic Gradient Descent will be presented and described.

2.4.2 Logistic Regression

The Logistic Regression fits into the linear probability models which consist on the application of the multiple regression model to a problem in which the dependent variable is binary (Wooldridge, 2013).

Therefore, given a binary classification problem- such as the probability prediction of the redemption of a promotional coupon- input $\boldsymbol{x}$ values are combined linearly using coefficient values $\boldsymbol{\beta}$ to predict an output value $\boldsymbol{y}$, the key difference from linear regression being that the output value to be modeled is a binary value (0 or 1) rather than a numeric value. In other words, the intention is to create a model in function of $\boldsymbol{x}$ with the conditional probability $P\ (y = 1 \mid x = xi)$. In the scope of this dissertation, the default class $\boldsymbol{y = 1}$ is the redemption of a coupon. To create this model, the Logit regression is defined by the following function:

$$log\frac{\mathrm{p(x)}}{1-\ \mathrm{p(x)}} = \beta 0 + \mathrm{x}\beta 1 \Leftrightarrow \hat{y} = \mathrm{p(x)} = \frac{e^{\beta 0+\mathrm{x}\beta 1}}{1+\ e^{\beta 0+\mathrm{x}\beta 1}} = \frac{1}{1+\ e^{-\beta 0+\mathrm{x}\beta 1}}$$

Where:

x, refers to the independent variable;

$\hat{y}$ is the predicted output: a real value between 0 and 1, that needs to be rounded to an integer value and mapped to a predicted class value.

β0, is the intercept parameter, also known as constant;

β1, represents the slope in the relation between x and y, *i.e.* it is the coefficient for the single input value *x1*;

e is the base of the natural logarithms (Euler's number).

Thus, $y = 1$ when $p \geq 0.5$ and $y = 0$ when $p < 0.5$. In other words, $y = 1$ whenever β0 + xβ1 is non-negative and $y = 0$, otherwise. The Logit regression thus presents a linear classifier (Shalizi, 2012).

2.4.3 Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is an the iterative process of **minimizing** a cost/ objective function $J(\theta)$ parameterized by a model's parameters $\theta \in \mathrm{R}d$ by updating the parameters in the opposite direction of the gradient of the objective function $\nabla\theta\ J(\theta)$, regarding the parameters.

Stochastic gradient descent (SGD) is defined by the following equation:

$$\theta = \theta - \eta \cdot \nabla\ \theta\, J\,(\theta;\, x\,(i);\, y\,(i))$$

In which, a parameter update is performed for each training example $x(i)$ and label $y(i)$, with a learning rate η, which determines the step size taken to attain the (local) minimum of the objective function.

Rephrasing, computing SGD involves knowing the form of the cost as well as the derivative so that, from a given point, the gradient follows the direction of the slope of the surface created by the objective function downhill (since the goal is cost minimization) until a valley is reached.

This optimization algorithm is used in machine learning to evaluate and update the coefficients θ at each iteration, to reduce the error for the next prediction, using the full training to converge to local optima.

3 Business Understanding

3.1 Resources

For the development of this dissertation, the Continent Card's Management Board allowed the access to its data history with sufficient variables, both at the promotional level as well as transactional, behavioral and demographic data of the client, to establish good basis for creation of the proposed model. In addition to the available data, it disposed of essential tools, namely SQL Developer and SAS Enterprise Guide.

3.2 Business Goals and Success Criteria

The Continente Card's Management Board ultimate motivation regarding this project is to use analytics as a tool to provide increased value to the customers and better services to the card's partners, while tackling the issue of customer loyalty and ensuring a greater likelihood of positive results.

To achieve it, the purpose of this work is to build an adaptive model that allows the prediction of the redemption rate of promotional mailing coupons for Continente. It is relevant to stress that throughout this dissertation the focus is solely placed on the Continente partner and on its promotional actions sent in the form of a coupon via letter to the client's home.

This knowledge will ultimately be used to predict the percentage of costumer adhesion to a certain promotional campaign and the redemption rate of a specific coupon; and is not focused in the redemption of a certain client individually.

However, the database will assemble both coupon and client data: demographic, transactional and behavioral data. This is the case on account of coupons not being sent to all customers evenly, and of the inexistence of defined rules of choice of customers to receive a coupon. Nonetheless, each coupon is defined according to a business strategy and must obey the optimization rules and space limitation of the mailing letter sent to the customers, as well as respect the priorities among Continente Card's partners and any

other rules that may arise during the process. Therefore, since customers are segmented in group to receive each coupon, the only possible way to predict the redemption rate of a coupon is by modeling the response and behavior of customers targeted to that coupon.

Using this knowledge, the business will be able to provide a better direct marketing promotional service to the partner and to plan better campaigns (Davenport, 2010). In addition, it will provide useful information to the partner for decision support regarding: stock management (given the percentage of customers that will engage with a given promotion, it is possible to manage the stock of the targeted products and prevent stock outs and shortage), strategic outline (to perceive and predict whether the objectives set for the campaign will be achieved or whether an adjustment is necessary in the promotion or in the target group) and sales forecast (it is possible to forecast the total amount of sales knowing the average basket of a certain group of customers and the redemption rate) (Granja, 2017).

On the other hand, it is also vital for the business to create an adaptive model. The process of automation will allow the model to perform self-adjustments in real time, tackling the issue of late response to market changes and the need for excessive human-driven efforts (Rico-Juan & Iñesta, 2014).

As for the success criteria, it is intended to obtain a redemption rate forecast with an error below the one of the current static model. It is also relevant for the business to perceive which variables directly affect the redemption of coupons, backed by analytical evidence rather than intuitive perception.

For the Continent Card's Management Board, it is important for this model to result in increased value for both the business of its partners and for the customer; and in case of success, to extend it to the other Continent Card partners of the ecosystem of brands and other channels of mailing.

3.3 Data Mining Goals and Success Criteria

The overall objective of this project is to obtain the predicted redemption rate of promotional mailing coupons for the Continente partner of the Continent Card, using an Adaptive Machine Learning structure.

However, more specific objectives are highlighted: (1) the use of Data Mining techniques in order to achieve better results than the current Logistic Regression static model, (2) pre-selection of attributes before applying the Data Mining techniques, to perceive their impact on the obtained results, (3) the creation of new metrics by employing the existing variables and data history, (4) the identification of the variables with greater relevance in this campaign, in order to potentiate the results of the next campaigns with similar molds, (5) to analyze the best methods for reducing the size of variables and infer whether they are effective or not, (6) study, analyze and adjust the use of Stochastic Gradient Descent on the current Logistic Regression model and compare the results and, finally, (7) deployment of the final model using the test set and analyze the results.

The data Mining success criteria is intertwined with the business one: it is intended to obtain a redemption rate forecast with an error below the error of the current static model. In addition, it is also crucial to identify the variables which are directly related to the redemption rate of coupons, as well to identify the ones which are not.

4 Data Extraction and Description

4.1 Data Collection

The process of data collection was executed using the SAS Enterprise Guide software.

The first task to be executed was to create a SAS program, to enable the extraction and construction of the data tables using SAS and SQL language, as the promotional, behavioral and demographic information is stored in tables of the Continent Card´s database, accessible through SQL oracle.

To provide a better understanding of the following steps, a summary scheme of the configuration of the databases used and their respective connections was made, and its illustration can be observed in figure 4.1.

4.2 Promotional Data

This process involved the gathering of promotional campaign information from Continente. In order to achieve this, SAS tables - which are created monthly by the operations teams - containing the information and parameterizations related to each campaign and its promotional coupons were consulted and gathered. These tables contain the information presented in table 4.1.

This information was collected for the 7 months, from July of 2017 to February 2018 (excluding September). The reasoning behind this has to do with Sonae's restructuration of data storage which made the inclusion of September's promotional data impossible in feasible time.

The information of 7 months of campaigns and respective coupons was compiled using SAS code, resulting on a table with 668 rows, that is, with 668 coupon offers with their respective parameterizations and their respective assigned segment codes.

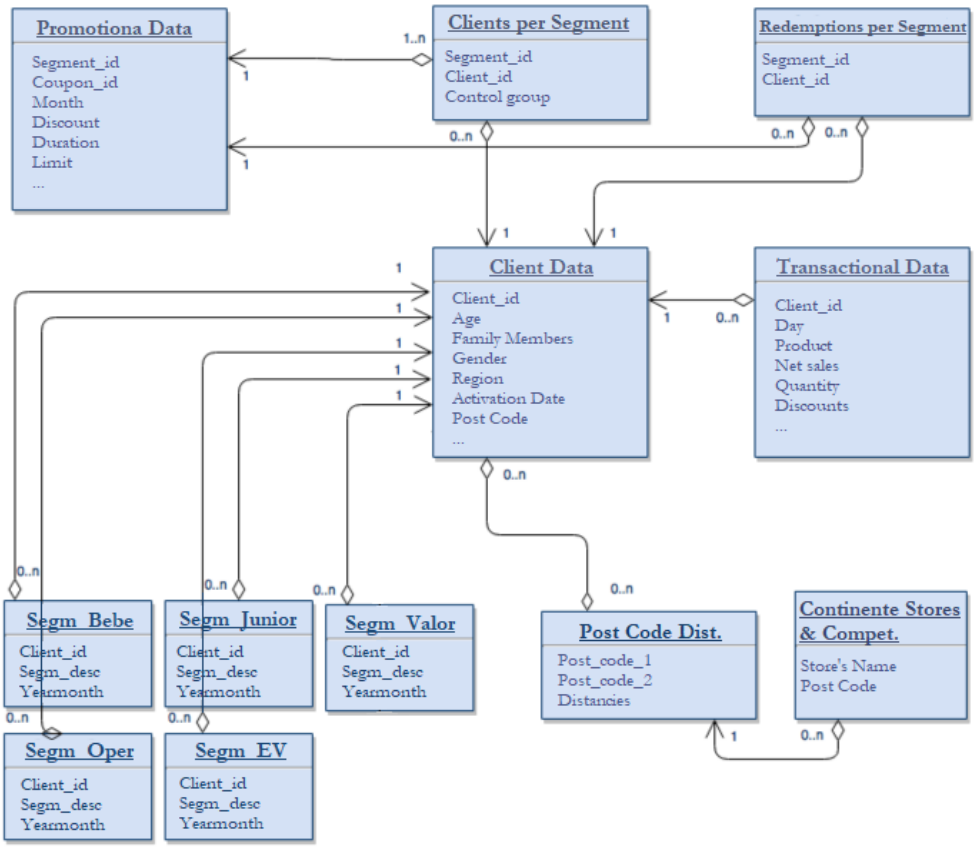

Figure 4.1 – Database scheme adapted from Granja 2017.

Variable	Description
Mailing_id	Identifies the Campaign
Segmento_id	8 digit code that identifies segmented clients
Promoção_id	10 digit code that identifies the Promotion
Data de Envio	Submission date of the Promotional Letter
Nr Dias Inicio	Number of days after the letter is received until the beginning of the promotion
Nr Dias	Duration of the promotional offer in days
Descrição	Brief description of the offer
Valor Desconto	Synopsis of the offer
Cupão_id	Digit code that identifies the coupon
Estratégia	Small acronym that identifies the coupon strategy
Mês	Campaign Month
Data_Ini	Start date of the offer
Data_Fim	End date of the offer
Valor Desconto	Discount value associated with the coupon
Limite Mínimo	Minimum purchase limit to obtain discount
Unidade Negocio	Business unit in which the discount is applied
Tipo de Desconto	Whether the discount is in the form of percentage or value

Table 4.1- Promotional data

4.2.1 Extraction of all the segmented Clients

The table Clients per Segment mentioned in figure 4.1, contains all customers which are registered for each segment of a promotional offer, meaning it contains all customers that have been selected to receive a coupon. This allows the extraction, through segment_id, of clients targeted to receive each promotional offer.

Using querys, a table with all the segmented clients and their respective promotions in the 7-month history was built by crossing the segment_id of Promotional Data table with the table of Clients per Segment. Thus, the initial structure of the database was obtained with 14135409 rows.

4.2.2 Extraction of all the Clients Who Redeemed Offers

In the Redemption per Segment table, a listing of all customers who have redeemed a mailing promotional campaign by segment_id is accessible. By crossing the list of segment_id with this table, the information regarding all the customers who effectively redeemed offers was obtained, that is, all the customers who used the promotional offers.

4.2.3 Construction of the response variable

Combining the resulting tables from steps 4.1.1 and 4.1.2, the response variable was created by crossing all clients assigned / targeted for each promotion with all customers who effectively redeemed coupons in the same segments / promotions. Therefore, the response variable is a binary variable in which the value 1 is assigned to customers who redeemed offers and 0 is attributed to the ones who did not. The resulting table was named Response_Variable.

4.3 Behavioral and Demographic Data

The process of gathering demographic and behavioral data was developed using a representative sample of 10% of clients, compiling the information of all past mailing campaigns. The reasoning behind this sample size had to do with the computational expenses associated with the massive database size, which could impact the ongoing processes of Sonae MC.

Having listed all clients targeted for promotions, the following step was to build a basis of behavioral variables for each client. Since each campaign is carried out on a monthly or bimonthly basis and the intention is to predict the behavior of the client in a given month, the goal is to analyze the targeted clients' past behavior in relation to that month.

Therefore, for each of the 7 months under study, a table containing the behavioral variables of the client in a period of one year prior to each respective month was created.

By accessing the tables containing transactional data, the following variables were extracted and created for each customer of Continente in the 12 months prior to the period of each campaign (U12M):

• Recency - How many days have passed since the customer has been to the Store or how long ago was the last visit to the store, in days.

• Nr_Trx_Tot - Total number of customer transactions in the last 12 months, in whole number.

• Vl_Tot - Total net sales to customers in the last 12 months, in euros.

• Qtd_Tot - Total amount of products purchased by the customer in the last 12 months, in real number.

• Num_prod_Tot- Distinct products purchased by the customer in the last 12 months, in whole number.

• Desconto_Tot - Total discounts obtained by the customer in the last 12 months, in euros.

• Num_Semanas_Tot - Total number of distinct weeks in which the client went to store in the last 12 months.

4.3.1 Extraction/ Construction of the Segment Value Variable

There is a customer segmentation made and calculated monthly for each Continente card partner based on the customer's behavior in each Insignia or Partner, which reflects the customer's value to the brand. They are, in the case of Continente, separated into categories with the names: Leal, Frequente, Ocasional and No Value. This segmentation is executed according to RFM criteria.

Therefore, the categories assigned to each customer in this segmentation in the 12 months prior to the month under analysis were extracted and, then, the mode of the classification was calculated for each customer.

Therefore, the most popular value segment of each client in the last 12 months was calculated for each month of campaign in analysis, resulting in 12 tables for each month with the following information:

- Customer ID - Customer identification number.
- Segm_Val_final - Classification of the segment Value of Continente's customers in the month prior to the promotional campaign.

4.3.2 Extraction/ Construction of the Distance Variable

Another calculated variable was the costumer's proximity to stores: both Continente or its competitors. Therefore, having access to the customer's, Continente & Partners and some competitors' post codes and geographical coordinates, the distances between each customer's address and the Continente Stores, as well as between each customer's address and some competitors were calculated.

• Customer Id - Customer identification number.

• Distance Continente Stores- Calculated distances, in km.

• Distance to Competitors - Calculated distances, in km.

4.3.3 Extraction/ Construction of the Coupon Rate of Participation in Promotional Offers

For each month, all campaigns for which the customers were selected were extracted as well as how many of the customers participated in those campaigns.
Using this information, the customer participation rate in campaigns is obtained which reveals which customers have a greater predisposition to use/ redeem future coupons. For the effect, 7 tables were created corresponding to each one of the 7months. For each month, the rate of mailing participation of each client is calculated in the year prior to each month.

• Customer ID - Client identification number.

• Nr_segmentos - Number of campaigns for which the client was selected and received a promotional offer.

• Nr_rebates - Number of redemptions / uses of promotional offers sent to the customer.

4.3.4 Extraction of Customer’s Demographics

For each client, several demographic variables were extracted, for instance:

• Idade - Client's age.

• Activation Date - Activation date of the Continente Card.

• Family Members - Number of household members.

• Gender - Man or Woman.

• Region and Islands – Region of Portugal where the customer lives: North, Center, South, Madeira, Azores.

4.3.5 Extraction of the Baby and Junior Segments

There is a segmentation, already created and updated monthly, in which various variables, whether provided by the customer or by purchasing habits, that distinguish clients as having babys, todlers or older children. For each of the 7 months of the campaign, the category of this segmentation in which each client belongs to was extracted, in the month prior to the campaign in analysis.

• Customer ID- Customer ID.

• Segm_baby - Category in Baby segmentation.

• Segm_junior - Category in the Junior segmentation.

4.3.6 Extraction of the Operational Segment

There is a segmentation, already created and updated monthly, in which various variables, whether provided by the customer or by purchasing habits, that distinguish customers into operational categories acordim to the RFM criteria: Loyal Large, Loyal Medium, Frequent Large, Frequent Medium, Occasional Small, Occasional Medium, Infrequent and No value. For each of the 7 months of the campaign, the category of this segmentation in which each client belongs to was drawn from the month prior to the campaign. This information was constructed with RFM and promotional data.

•CustomerID - Customer ID.

Segm_Oper - Category in operational segmentation.

4.3.7 Extraction of the Lifestyle Segment

There is a segmentation, already created and updated monthly, in which several variables, whether provided by the customer or by purchasing habits throughout the ecosystem of the Continent Card, that distinguishes customers into several categories which are meant to be different lifestyles: Saudáveis Exigentes, Urbanos Sofiticados, Generalistas Disciplinados, Práticos Paternalistas, Tradicionais Frequentes, Económicos Focados and Promocionais atentos . For each of the 7 months of campaign, the category of this segmentation for each customer was extracted, in the month prior to the campaign.

•CustomerID - Customer Identification number.
• Segm_Oper - Category in the Lifestyle segmentation.

4.4 Data Description

4.4.1 Extracted Variables

As a way of summarizing all the extracted variables mentioned in the previous sections, a table was built containing the description of each variable. This description can be observed in table 4.2.

Variable	Description	Type
Açores	Flags a promotional offer designated for the Azores region.	Binary Variable
Bebe	Flags a promotional offer related to baby products.	Binary Variable
DATA_ATIVACAO	Date of adhesion to the Continent Card, in days.	Discrete numerical variable.
DESCONTO_TOT	Total customer discounts in the U12M.	Continuous numerical variable.
EXCLUSAO	Flag that indicates whether a costumer belongs to a control group=1 or not=0.	Binary Variable
FAMILY_MEMBERS	Number of household members.	Discrete numerical variable.
GENERO	Indicates the gender of the client.	Nominal categorical variable.
IDADE	Age of the client.	Discrete numerical variable.
LIMITE	The variable is 1 if there is a minimum purchase limit to obtain a discount, otherwise it is 0.	Binary variable
Madeira	Flags a promotional offer designated for the Madeira region.	Binary Variable
Mês	Indicates the Campaign Month.	Nominal categorical variable.
N_DIAS	Duration of the promotional offer in days	Numerical descrete variable
N_DIAS_INI_MAIL	Number of days after the letter is received, until the beginning of the promotion.	Numerical descrete variable
N_SEMANAS_TOT	Total number of weeks in which the customer went to Continente on U12M.	Continuous numerical variable.
natal	Equals 1 when it is a promotional offer from the Christmas period and 0 in the rest of the year.	Binary Variable
nr_campanhas	Number of promotional offers the customer received, in which he belonged to the target group.	Discrete numerical variable
nr_rebates	Number of times the customer used promotional coupons.	Discrete numerical variable
NUM_PROD	Total of different products purchased by the customer in U12M.	Continuous numerical variable.
NUM_TRX	Total number transactions by customer.	Continuous numerical variable.
NUM_TRX_TOT	Total number of customer transactions in the last 12 months prior to the campaign month.	Continuous numerical variable.
per_desconto	Discount percentage of the promotional offer.	Numeric variable in%.
PROD__TOT	Total of different products purchased in U12M by the customer.	Continuous numerical variable.
QTD_TOT	Total quantity of products purchased on U12M by the customer.	Continuous numerical variable.
REBATE	This is the target variable. It is equal to 1 when there was a coupon redemption and 0 when there was not.	Binary Response Variable
RECENCY	Number of days spent since the customer last visited the store.	Numerical descrete variable
REGIAO	Region of the country where the customer lives.	Nominal categorical variable.
REGRESSO_AULAS	Flag which indicates if the coupon correponds to a "back to school" offer. It will be 1 when it is and 0 when it is not.	Binary variable
SEGM_BEBE	Baby segmentation category in which the client belongs in the month prior to the campaign.	Nominal categorical variable.
SEGM_EV	Lifestyle category the customer belongs to in the month prior to the campaign.	Nominal categorical variable.
SEGM_JUNIOR	Junior segmentation category in which the customer belongs to in the month prior to the campaign.	Nominal categorical variable.
SEGM_OPER	Operational category in which the customer belongs to in the month prior to the campaign.	Nominal categorical variable.
SEGM_VALOR	Mode of the Continente segmentation category value in which the client belonged in the U12M prior to the campaign.	Nominal categorical variable.
UNIDADE_NEGOCIO	Business unit in which the discount is applied. Each category will be transformed into a a binary variable.	Nominal categorical variable
VL_TOT	Total net sales in the U12M.	Continuous numerical variable.

Table 4.2- Descriptive summary of all the Extracted Variables

4.4.2 Calculated Metrics

At this stage, the metrics calculated out of existing variables will be explained. The advantage of creating new metrics is to emphasize certain aspects of customer behavior that may result in gains in the forecast of the probability of a customer redeeming a promotional offer and consequently of the general redemption rate that we intend to predict.

The following created metrics are identified, described and explained in 2 tables: 4.3 contains the binary variables built (most concern product category) and 4.4 contains the numeric variables:

Variable	Description
flag_comp_p1	Flag which indicates a recent purchase. It equals 1 when the recency variable is less than 120 days and 0 when it is higher. If recency <120 days, Flag_comp_p1=1, otherwise it is 0.
flag_comp_tot_p1	Flag which indicates the total of recent purchases. It will be 1 when the total recency variable is less than 120 days and 0 when it is higher. If recency_tot <120 days, Flag_comp_tot_p1=1, otherwise it is 0.
MERCEARIA SALGADA	Flag which indicates the business unit of the pomotional offer: it will be 1 if it belongs to MERCEARIA SALGADA, and 0 otherwise.
MERCEARIA DOCE	Flag which indicates the business unit of the pomotional offer: it will be 1 if it belongs to MERCEARIA DOCE, and 0 otherwise.
BEBIDAS	Flag which indicates the business unit of the pomotional offer: it will be 1 if it belongs to BEBIDAS, and 0 otherwise.
HIGIENE E BELEZA	Flag which indicates the business unit of the pomotional offer: it will be 1 if it belongs to HIGIENE E BELEZA, and 0 otherwise.
LIMPEZA DO LAR	Flag which indicates the business unit of the pomotional offer: it will be 1 if it belongs to LIMPEZA DO LAR, and 0 otherwise.
CONGELADOS	Flag which indicates the business unit of the pomotional offer: it will be 1 if it belongs to CONGELADOS, and 0 otherwise.
LACTICINIOS	Flag which indicates the business unit of the pomotional offer: it will be 1 if it belongs to LATICINIOS, and 0 otherwise.
TALHO	Flag which indicates the business unit of the pomotional offer: it will be 1 if it belongs to TALHO, and 0 otherwise.
PEIXARIA	Flag which indicates the business unit of the pomotional offer: it will be 1 if it belongs to PEIXARIA, and 0 otherwise.
CHARCUTARIA&QUEIJOS	Flag which indicates the business unit of the pomotional offer: it will be 1 if it belongs to CHARCUTARIA&QUEIJOS, and 0 otherwise.
CULTURA	Flag which indicates the business unit of the pomotional offer: it will be 1 if it belongs to CULTURA, and 0 otherwise.
FRUTAS E LEGUMES	FRUTAS E LEGUMES, and 0 otherwise.
PADARIA	Flag which indicates the business unit of the pomotional offer: it will be 1 if it belongs to PADARIA, and 0 otherwise.
TAKEAWAY	Flag which indicates the business unit of the pomotional offer: it will be 1 if it belongs to TAKEAWAY, and 0 otherwise.
LAZER	Flag which indicates the business unit of the pomotional offer: it will be 1 if it belongs to LAZER, and 0 otherwise.
CASA	Flag which indicates the business unit of the pomotional offer: it will be 1 if it belongs to CASA, and 0 otherwise.
BRICO & AUTO	Flag which indicates the business unit of the pomotional offer: it will be 1 if it belongs to BRICO&AUTO, and 0 otherwise.
PET&PLANTS	Flag which indicates the business unit of the pomotional offer: it will be 1 if it belongs to PET&PLANTS, and 0 otherwise.
TEXTIL	Flag which indicates the business unit of the pomotional offer: it will be 1 if it belongs to TEXTIL, and 0 otherwise.

Table 4.3- Descriptive summary of all the calculated Binary Variables

Variable	Description	Computation
per_descontos	Percentage of discounts conceded in the U12M.	Emited discounts/Total Discounts
cesta_med	Average shopping basket of the customer in the U12M, in Euros.	Net sales/number of transactions
cesta_med_tot	Average shopping basket in the U12M, in Euros.	Total net sales/ Total number of transactions
desconto_med	Average discount per transaction, in Euros.	Emited discounts/Number of transactions
desconto_med_tot	Total average discount per total number of transactions, in Euros.	Total Discounts/ Total number of transactions
desconto_vend	Discount divided by net sales, in percentage.	(Emited discounts/ Net sales)* 100
desconto_vend_tot	Total discount divided by total net sales, in percentage.	(Total discounts/ Total Net sales)* 100
dist_concorr	Distance from customer's adress to the competitor, in kms.	Min Dist (Lidl, Minipreço, Intermarche, Pingo Doce, Jumbo)
dist_continente	Distance from the customer's adress to Continente stores, in Kms.	Min Dist (Continente Bom Dia, Continente, Continente Modelo)
expetativa_desconto	Expected discount. It will be the discount percentage multiplied by the average basket of the customer, in euros.	discount percentage * the average basket
index_conco	Competition Index. Percentage indicating that the larger it is, the greater the proximity of competition in the proximity of the store.	sum (dist continente_competitors/n) *100
num_prod_med	Average number of product purchasedby transaction, in euros.	nr of products/nr of transactions
per_produtos	Percentage of products sold in the U12M.	(nr of products/ total nr of products)*100
per_quant	Percentage of quantities sold in the U12M.	(quantity/ total quantity)*100
per_semanas	Percentage of weeks the customer went to the store on U12M, in percentage.	(nr weeks/ total nr weeks)*100
PER_SEMANAS_tot	Percentage of of total weeks the customer went to the store on U12M, in percentage.	(total number of weeks / 54)*100
per_trx	Percentage of transactions in the U12M.	(nr transactions/ totalnr transactions)*100
per_vl	Percentage of net sales in the U12M.	(net sales/ total netsales)*100
preco_med_total	Average cost per product purchased. Purchases divided by the quantity of products, in euros.	net sales/quantities
preco_med_tots	Total average cost per product purchased. Total purchases divided by the quantity of products, in euros.	total net sales/total quantities
PROD_SEM_tot	Quantity of different products purchased per week. Total of discrete items bought,split by the number of weeks the store was visited , in euros.	total nr of products/total nr of weeks
qm_tot	Average quantitypurchased by number of transactions.	Quantity/ number of transactions
QTD_SEM_tot	Quantity of products purchased per week. Total amount of items purchased divided by the number of weeks the store was visited , in euros.	total quantity/ total nr of weeks
RECENCY_TOT	Maximum value of recency, measured in days.	= 365
taxa_reb_mail_cli	Rate of redemption of promotional offers on U12M. Total offers used to split by total offers received and targeted.	number of redemptions/Number of segments
TRX_SEM_tot	Number of transactions per week in U12M.	Total number of transactions/Total number of weeks
var_dist	Distance from customer's adress to the competitors divided by customer distance to Continente stores.	dist_competitors/dist_Continente
var_recency	Variation in product group recency versus total store recency.	Recency/ Total Recency
var_semana	Variation in number of transactions related to number of weeks.	Number of transactions/Number of weeks

Table 4.4 Descriptive summary of all the calculated Numeric Variables

4.5 Data Exploration

At this stage the variables will be analyzed. The first analysis performed on the variables was to calculate the measures of Location and Dispersion. The results are presented in table 4.5:

Variable	Mean	Median	Std Dev	Min	Max	Coef. of Variation
N_DIAS	13.2444499	14.0000000	1.6647663	7.0000000	14.0000000	12.5695390
N_DIAS_INI_MAIL	14.2409248	14.0000000	13.3158254	0	28.0000000	93.5039375
PER_DESCONTO	0.2757125	0.2500000	0.0639538	0.1500000	0.5000000	23.1958373
RECENCY	256.6952583	292.0000000	100.1653191	0	364.0000000	39.0211022
NUM_TRX	6.7247231	3.0000000	9.8236810	1.0000000	674.0000000	146.0830545
NUM_PROD	5.1642177	3.0000000	6.8744333	1.0000000	200.0000000	133.1166443
RECENCY_TOT	10.8123236	5.0000000	13.8694905	0	364.0000000	128.2748373
NUM_TRX_TOT	53.4915203	39.0000000	53.0017658	0	1214.00	99.0844261
IDADE	49.1606425	47.7000000	16.3931621	18.0000000	100.0000000	33.3461103
DATA_ATIVACAO	8.1207413	9.7000000	3.0409047	0.2000000	11.1000000	37.4461472
FAMILY_MEMBERS	3.2213306	3.0000000	4.9866985	0	99.0000000	154.8024446
var_recency	64.0875960	24.5700000	90.8685336	0	364.0000000	141.7880202
perc_semanas	11.1829080	5.6000000	13.8013493	1.9000000	98.1000000	123.4146724
var_semana	1.0527534	1.0000000	0.1524779	1.0000000	14.7000000	14.4837218
per_trx	16.1351031	11.0000000	16.0634117	0	100.0000000	99.5556805
per_vl	2.6169809	1.0000000	4.9663718	0	100.0000000	189.7748593
per_descontos	2.0161704	0	9.4795899	0	100.0000000	470.1780148
per_produtos	2.2974155	1.0000000	3.0215503	0	100.0000000	131.5195370
per_quant	2.3416495	1.0000000	3.9402059	0	100.0000000	168.2662497
per_semanas	0.9998550	1.0000000	0.0120418	0	1.0000000	1.2043543
cesta_med	3.9597712	2.3700000	6.4322565	0.0100000	1261.87	162.4401042
qm_tot	1.8643441	1.4000000	2.2491936	0	3082.00	120.6426232
preco_med_tot	2.3345487	1.5000000	10.0792516	0	14218.30	431.7430399
desconto_vend	0.8704572	0	3.8712743	0	625.0000000	444.7403235
desconto_med	0.0695466	0	0.9696513	0	468.1600000	1394.25
num_prod_med	0.9937861	1.0000000	0.5936979	0	52.0000000	59.7410116
cesta_med_tot	27.5813949	22.6300000	19.6767421	0	917.6600000	71.3406345
preco_med_tots	1.6525344	1.5800000	0.5442381	0	324.3900000	32.9335433
PER_SEMANAS_tot	0.5208534	0.5200000	0.2685652	0	0.9800000	51.5625207
TRX_SEM_tot	1.6727903	1.4300000	0.8438384	0	36.5000000	50.4449613
desconto_vend_tot	1.0068010	0.4000000	1.8803504	0	617.1000000	186.7648617
desconto_med_tot	0.2995167	0.1000000	0.6761240	0	121.3000000	225.7383694
QTD_SEM_tot	26.4043285	22.5400000	17.1386800	0	711.5100000	64.9086002
PROD_SEM_tot	10.8881523	10.0000000	5.4947482	0	140.0000000	50.4653870
expetativa_desconto	1.1062120	0.6300000	1.7803011	0	332.4200000	160.9367020
taxa_reb_mail_cli	0.1166418	0.0590000	0.1488764	0	1.0000000	127.6355603
dist_continente	4.1376314	2.1299872	5.3787906	0.0089328	249.8878491	129.9968543
dist_concorr	10.1373171	1.0773831	88.0599952	0.0021756	1466.95	868.6716025
var_dist	2.3355975	0.6212858	21.3204120	0.000131454	9225.31	912.8461616
index_conco	53.1627023	60.0000000	35.7585488	0	100.0000000	67.2624741

Table 4.5- Location and Dispersion Measures

Some conclusions can be drawn from the values obtained. The presence of severe outliers is noticeable in several instances, by comparing the maximum and mean values. Variables that present these characteristics are, for example: Number of Transactions, Number of Products, Distances and Discounts.

There is a frankly positive asymmetry for some variables, such as Dist_Continente, Dist_Concorrente, Qds_Sem_tot, and Var_Distan, Var_Recency. Negative asymmetry distributions are verified for the variables Data_Ativacao, Index_Conc, Recency.

After an initial analysis of the measurements of location, dispersion and shape it is intended to graphically explore all variables.

For the numerical variables the boxplots provide a visual aspect to the analysis already made using table 4.5. The boxplots of all the numeric variables can be seen in Figures 4.2, 4.3, 4.4, 4.5.

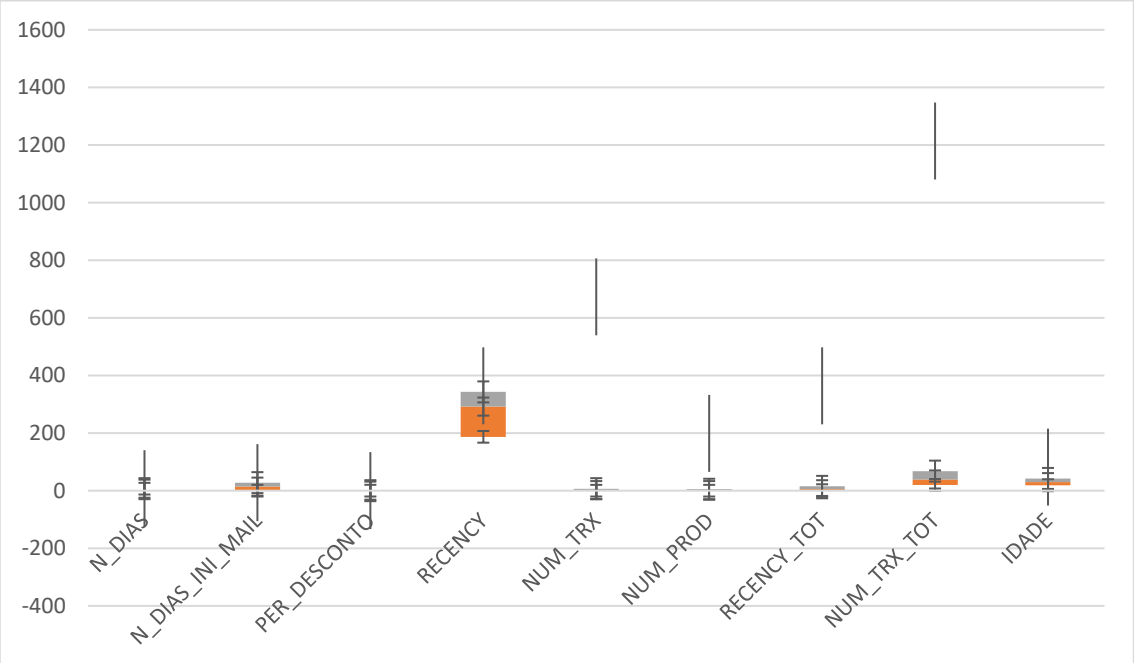

Figure 4.2- Box Plots I- Numerical Variables.

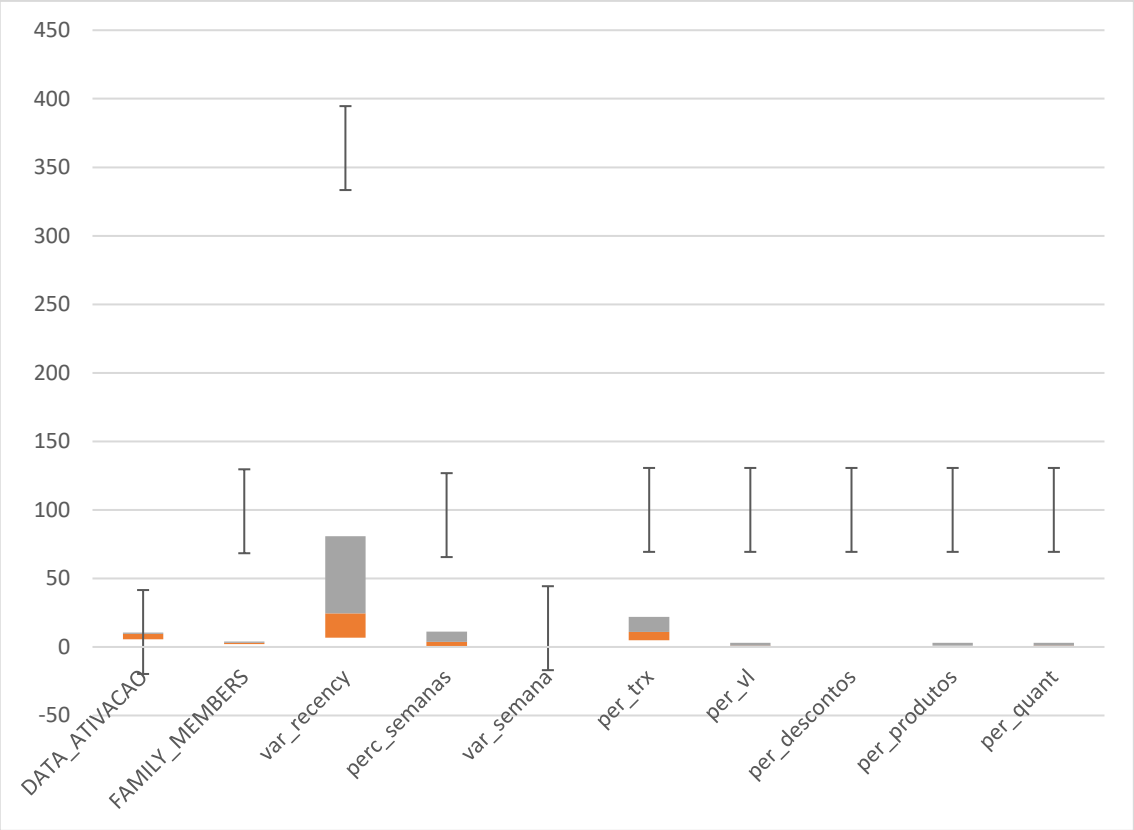

Figure 4.3 - Box Plots II- Numerical Variables.

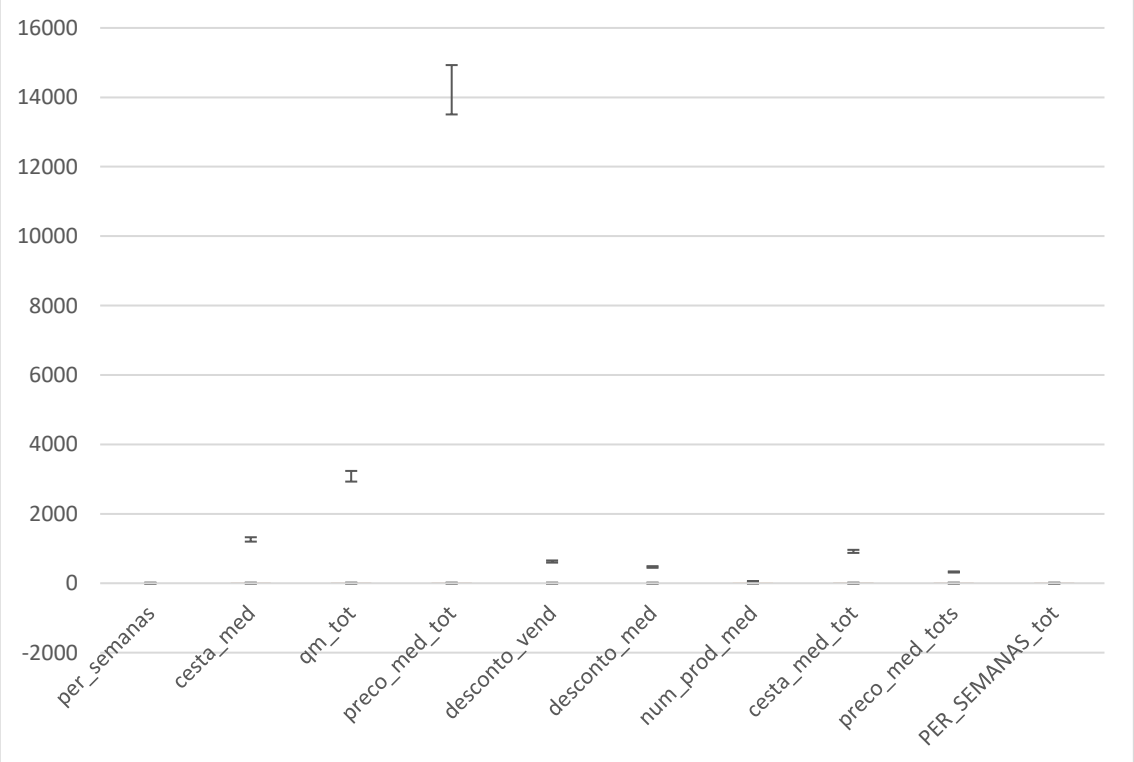

Figure 4.4- Box Plots III- Numerical Variables.

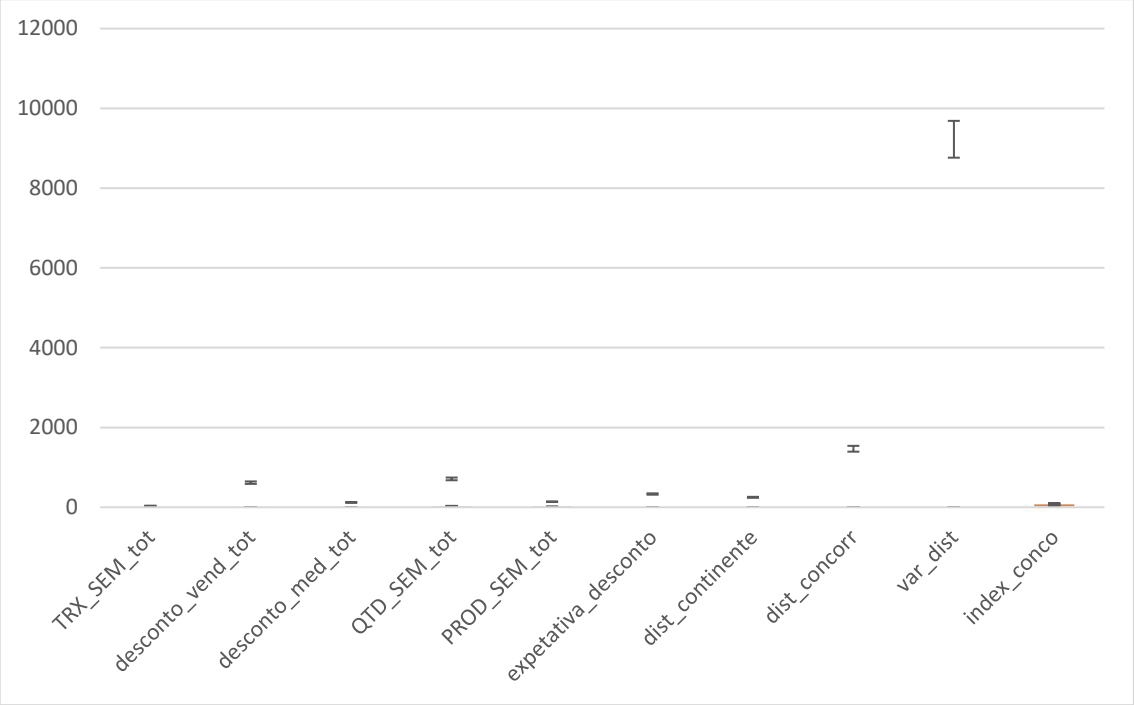

Figure 4.5- Box Plots IV- Numerical Variables.

The histograms of the categorical variables can be seen in Figure 4.6. By observing it some conclusions can be drawn:

- There are more women than men in the database.
- The Redemption Response Variable is unbalanced, that is, about 96% are observations with redemption equal 0 and only 4% with redemption equal to 1.
- The distribution of observations by month of the year is uniform, though December and November are a little more representative, perhaps because of the existence of more promotional campaigns or there are more targeted customers in these months.
- In the region distribution, the South and North are the most representative and Madeira and Azores are the least.
- In the Segm_value segmentation the majority of clients are classified as Frequent Large, Occasional Large and Frequent Medium.
- In the Business Unit classification, the most represented category is Mercearia Salgada and the least one is Textil.
- 13% of the Clients had made a recent purchase.
- It was also reported that around 13% of coupons corresponded to a Christmas offer.

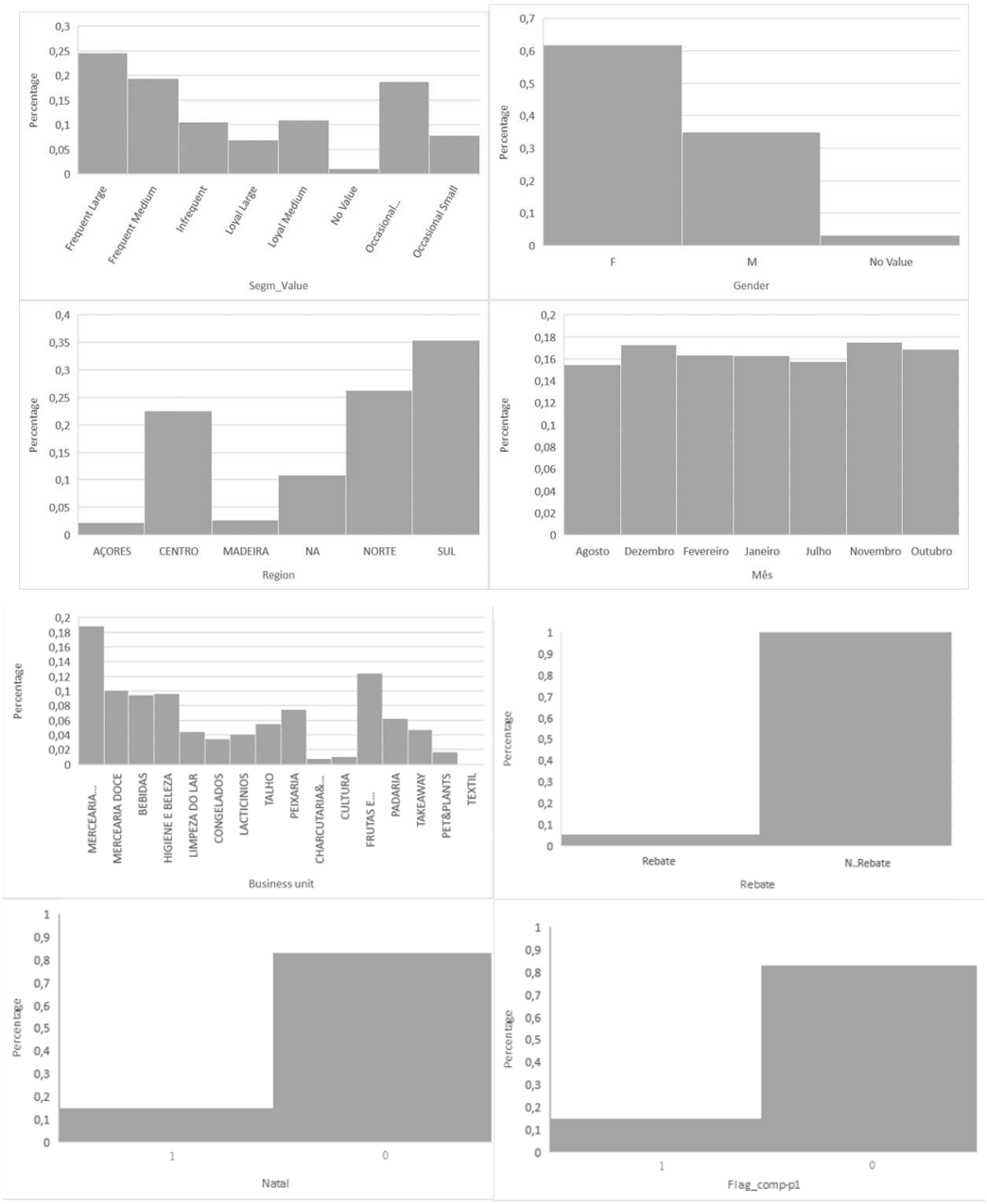

Figure 4.6- Histograms- Categorical Variables

Some variables were not represented: there were no clients belonging to Segm_baby and the Segm_Junior segmentations. In addition, the “Back to school” promotional offers and the variable Exclusão flagged less than 1% of the cases.

4.6 Data Correlation

Before the modelling phase it is important to analyze the correlation between explanatory variables. Correlation can be analysed using three coefficients: Pearson, Spearman and Phi. As what is intended is to calculate the correlation between the quantitative variables, the Pearson coefficients are the most adequate. For ordinal variables the Spearman coefficients are preferable and for the nominal ones, Phi should be used (Anderson,1958).

REBATE	N_DIAS	N_DIAS_INI_MAIL	PER_DESCONTO	LIMITE	BEBE	REGRESSO_AULAS	MERCEARIA SALGADA	MERCEARIA DOCE	BEBIDAS
1	0,020200649	0,04062886	-0,031149143	-0,03155751	-0,012882399	-0,005191724	0,031496229	0,057169573	-0,025301164
0,020200649	1	0,485378057	-0,034746189	-0,039441658	0,001502823	0,045572948	0,018762672	0,066376553	0,003250174
0,04062886	0,485378057	1	-0,034731535	0,066797494	-0,051811919	0,103757207	-0,005351213	0,038357	0,066330841
-0,031149143	-0,034746189	-0,034731535	1	-0,119439787	-0,025943647	0,096242829	-0,065374435	-0,078156241	0,308494912
-0,03155751	-0,039441658	0,066797494	-0,119439787	1	-0,015326467	0,422773298	0,009014939	0,062208856	0,125260077
-0,012882399	0,001502823	-0,051811919	-0,025943647	-0,015326467	1	-0,006479621	-0,030911722	-0,02151352	-0,020737298
-0,005191724	0,045572948	0,103757207	0,096242829	0,422773298	-0,006479621	1	-0,048102517	-0,033477736	-0,032269836
0,031496229	0,018762672	-0,005351213	-0,065374435	0,009014939	-0,030911722	-0,048102517	1	-0,159709104	-0,153946689
0,057169573	0,066376553	0,038357	-0,078156241	0,062208856	-0,02151352	-0,033477736	-0,159709104	1	-0,107141723
-0,025301164	0,003250174	0,066330841	0,308494912	0,125260077	-0,020737298	-0,032269836	-0,153946689	-0,107141723	1

HIGIENE E BELEZA	LIMPEZA DO LAR	CONGELADOS	LACTICINIOS	TALHO	PEIXARIA	·ARCUTARIA&QUEIJ(	CULTURA	FRUTAS E LEGUMES	PADARIA
1	-0,070256211	-0,061560199	-0,06658369	-0,078076203	-0,092428122	-0,029249194	-0,032712619	-0,121924531	-0,083882579
-0,070256211	1	-0,040751944	-0,04407742	-0,051685294	-0,061186053	-0,019362535	-0,021655271	-0,080712241	-0,05552903
-0,061560199	-0,040751944	1	-0,038621707	-0,045287911	-0,053612706	-0,016965924	-0,018974875	-0,070722026	-0,048655885
-0,06658369	-0,04407742	-0,038621707	1	-0,048983535	-0,057987658	-0,018350392	-0,020523279	-0,076493149	-0,052626346
-0,078076203	-0,051685294	-0,045287911	-0,048983535	1	-0,067996475	-0,021517716	-0,024065649	-0,089696059	-0,061709786
-0,092428122	-0,061186053	-0,053612706	-0,057987658	-0,067996475	1	-0,025473089	-0,028489382	-0,106183933	-0,073053241
-0,029249194	-0,019362535	-0,016965924	-0,018350392	-0,021517716	-0,025473089	1	-0,009015562	-0,033602267	-0,023117947
-0,032712619	-0,021655271	-0,018974875	-0,020523279	-0,024065649	-0,028489382	-0,009015562	1	-0,037581143	-0,025855365
-0,121924531	-0,080712241	-0,070722026	-0,076493149	-0,089696059	-0,106183933	-0,033602267	-0,037581143	1	-0,096366581
-0,083882579	-0,05552903	-0,048655885	-0,052626346	-0,061709786	-0,073053241	-0,023117947	-0,025855365	-0,096366581	1

PADARIA	TAKEAWAY	PET&PLANTS	TEXTIL	EXCLUSAO	RECENCY	NUM_TRX	NUM_PROD	RECENCY_TOT	NUM_TRX_TOT	IDADE
-0,057128612	1	-0,028573757	-0,0065362	-0,000493737	0,002858413	-0,006401797	-0,015473822	-0,022655417	0,025053438	-0,023172572
-0,03316048	-0,028573757	1	-0,003794	-0,000286591	0,003234721	0,006223453	0,005763221	-0,021980945	0,030317502	-0,006175562
-0,007585424	-0,006536216	-0,003793967	1	0,001014504	-0,009511988	-0,012147978	-0,011910336	0,005724803	-0,007767797	-0,002194523
-0,000572993	-0,000493737	-0,000286591	0,0010145	1	0,000575199	0,000845497	0,001671764	-0,000593344	0,000688328	-0,000632091
-0,002987129	0,002858413	0,003234721	-0,009512	0,000575199	1	0,410111967	0,373019908	-0,138325448	0,22786059	0,048027997
0,012751355	-0,006401797	0,006223453	-0,012148	0,000845497	0,410111967	1	0,690824001	-0,206201753	0,425738342	0,020769628
-0,012951734	-0,015473822	0,005763221	-0,0119103	0,001671764	0,373019908	0,690824001	1	-0,146779176	0,232262542	-0,021860494
0,023125569	-0,022655417	-0,021980945	0,0057248	-0,000593344	-0,138325448	-0,206201753	-0,146779176	1	-0,364946184	0,004682725
-0,029253347	0,025053438	0,030317502	-0,0077678	0,000688328	0,22786059	0,425738342	0,232262542	-0,364946184	1	0,006421765
-0,000927844	-0,023172572	-0,006175562	-0,0021945	-0,000632091	0,048027997	0,020769628	-0,021860494	0,004682725	0,006421765	1

DATA_ATIVACAO	FAMILY_MEMBERS	MADEIRA	ACORES	var_recency	flag_comp_p1	flag_comp_tot_p1	perc_semanas	var_semana	per_trx
1	0,026672703	-0,005195727	0,019505848	0,086849838	-0,114092818	0,005575287	0,115450028	0,001876612	-0,050363198
0,026672703	1	0,015343528	0,006247307	0,002960792	-0,006442561	-4,9586E-06	0,007181613	0,003142535	-0,000432797
-0,005195727	0,015343528	1	-0,016425148	0,029930711	-0,018159802	0,000738534	0,053450837	0,038793232	-0,022122613
0,019505848	0,006247307	-0,016425148	1	0,009005461	-0,012063127	0,001542795	0,015311727	0,005424788	-0,008658044
0,086849838	0,002960792	0,029930711	0,009005461	1	-0,216411203	0,014028887	0,259746311	0,065621733	0,021120272
-0,114092818	-0,006442561	-0,018159802	-0,012063127	-0,216411203	1	0,008037208	-0,239013445	-0,048891726	-0,098071567
0,005575287	-4,9586E-06	0,000738534	0,001542795	0,014028887	0,008037208	1	0,012507497	0,002710078	-0,02824664
0,115450028	0,007181613	0,053450837	0,015311727	0,259746311	-0,239013445	0,012507497	1	0,34158363	0,529066815
0,001876612	0,003142535	0,038793232	0,005424788	0,065621733	-0,048891726	0,002710078	0,34158363	1	0,185885018
-0,050363198	-0,000432797	-0,022122613	-0,008658044	0,021120272	-0,098071567	-0,02824664	0,529066815	0,185885018	1

per_vl	per_descontos	per_produtos	per_quant	per_semanas	cesta_med	qm_tot	preco_med_tot	desconto_vend	desconto_med	num_prod_med
1	0,21950947	0,65520098	0,749003214	0,006346227	0,437860232	0,246710222	0,062126462	0,034792241	0,090876047	0,095544126
0,21950947	1	0,114409278	0,144804543	0,002561486	0,202865053	0,10704024	0,026720165	0,503520185	0,332747442	0,006610715
0,65520098	0,114409278	1	0,783956328	0,009157227	0,078387506	0,193806263	-0,021942375	-0,012962608	0,000386073	0,343579271
0,749003214	0,144804543	0,783956328	1	0,007157432	0,173540836	0,386177569	-0,025351434	-0,001888277	0,015812596	0,141518323
0,006346227	0,002561486	0,009157227	0,007157432	1	0,00059334	4,35821E-05	4,34966E-05	0,000354256	0,000336992	-0,002262378
0,437860232	0,202865053	0,078387506	0,173540836	0,00059334	1	0,358107489	0,183783115	0,140947675	0,385779351	0,172475587
0,246710222	0,10704024	0,193806263	0,386177569	4,35821E-05	0,358107489	1	-0,017365332	0,057875687	0,109823671	0,250530468
0,062126462	0,026720165	-0,021942375	-0,025351434	4,34966E-05	0,183783115	-0,017365332	1	0,023760822	0,044026528	-0,000967296
0,034792241	0,503520185	-0,012962608	-0,001888277	0,000354256	0,140947675	0,057875687	0,023760822	1	0,464714496	0,017144637
0,090876047	0,332747442	0,000386073	0,015812596	0,000336992	0,385779351	0,109823671	0,044026528	0,464714496	1	0,073668271
0,095544126	0,006610715	0,343579271	0,141518323	-0,002262378	0,172475587	0,250530468	-0,000967296	0,017144637	0,073668271	1

cesta_med_tot	preco_med_tots	PER_SEMANAS_tot	TRX_SEM_tot	desconto_vend_tot	desconto_med_tot	QTD_SEM_tot	PROD_SEM_tot	expetativa_desconto	natal
1	0,303765264	-0,064867606	-0,224951442	0,058733914	0,351880768	0,657246049	0,679902623	0,163477034	-0,002183413
0,303765264	1	-0,200033185	-0,162037013	0,074728908	0,154350015	-0,132353209	0,034487332	0,147998224	0,022047126
-0,064867606	-0,200033185	1	0,525053354	0,078168654	0,040079011	0,338717143	-0,031450738	-0,018822481	-0,032624978
-0,224951442	-0,162037013	0,525053354	1	-0,00960962	-0,075370618	0,342750155	0,108597276	-0,030104327	-0,016514224
0,058733914	0,074728908	0,078168654	-0,00960962	1	0,755026391	0,030806621	-0,008378101	0,051409471	-0,01347779
0,351880768	0,154350015	0,040079011	-0,075370618	0,755026391	1	0,218747851	0,17165144	0,103620705	-0,012262431
0,657246049	-0,132353209	0,338717143	0,342750155	0,030806621	0,218747851	1	0,700227245	0,085519311	-0,018695141
0,679902623	0,034487332	-0,031450738	0,108597276	-0,008378101	0,17165144	0,700227245	1	0,057034022	-0,011024987
0,163477034	0,147998224	-0,018822481	-0,030104327	0,051409471	0,103620705	0,085519311	0,057034022	1	0,065925163
-0,002183413	0,022047126	-0,032624978	-0,016514224	-0,01347779	-0,012262431	-0,018695141	-0,011024987	0,065925163	1
0,15947825	-0,111506432	0,4615917	0,166346514	0,188554515	0,190581714	0,339492626	0,124547081	0,027287874	0,068202612
0,091921458	0,052920365	-0,188698767	-0,111032276	-0,025557937	0,000927702	0,001341123	0,064498777	0,021661809	0,013837319
0,008888087	0,017644585	0,028157999	0,021932098	0,007503209	0,006130905	0,014869557	-0,001346141	-0,001889843	0,076597812
-0,000752127	0,008474845	0,036351217	0,030442996	0,006513285	0,002946948	0,013019059	-0,004680397	-0,002933501	0,061584533
-0,057891627	-0,014247733	0,085506048	0,050647202	0,045082199	0,025047528	-0,027141088	-0,032238448	-0,016172659	-0,020292086

taxa_reb_mail_cli	dist_continente	dist_concorr	var_dist	index_conco
1	-0,082931762	-0,026745942	-0,017469963	0,070012079
-0,082931762	1	0,071519633	0,00412874	-0,527925536
-0,026745942	0,071519633	1	0,786209794	-0,155918279
-0,017469963	0,00412874	0,786209794	1	-0,113328983
0,070012079	-0,527925536	-0,155918279	-0,113328983	1

Table 4.6- Spearman Coefficients

There are several positive and negative correlated variables. Some of these correlations were expected since most of the metrics were built from other existing variables.

5 Data Preparation

In the previous phase of data exploration, some variables initially considered were already excluded. In this chapter data cleaning, treatment and preparation is discussed.

Finally, a table was created for each campaign month, containing all the variables extracted from the company's databases as well as the calculated metrics. Subsequently, they were gathered in a final table per client with 14135409 rows. This chapter contains the steps of data cleaning and preparation the database suffered before the modelling stage.

5.1 Data Cleaning

5.1.1 Missing Value Treatment

The treatment of existent missing values is essential for the proper functioning of the logistic regression. In this subsection the techniques used to deal with the missing values are explored.

Treatment of missing values detected in categorical variables:

• Target value - missing value replaced by 'No Value'.

• Baby Segmentation - missing value replaced by 'No Baby'.

• Junior Segmentation - missing value replaced by 'No Junior'.

• Region - missing value replaced by 'NA'.

• Gender - missing value replaced by 'NA'.

Some incongruent values were also found in the data, which must be corrected. For example, the variable total discounts cannot assume a negative value. In these situations, the corrections were performed depending on the variable:

• If Recency <0, then Recency = 0. If for some reason recency is less than 0, then it will be replaced by 0.

• If Total Recency <0, then Recency_tot= 0. If for some reason the total recency is less than 0, then it will be replaced by 0.

• If the Number of Transactions <0 then num_trx = 0. If for some reason the number of transactions is less than 0, then it will be replaced by 0.

• If the Total Number of Transactions <0 then num_trx_tot = 0. If for some reason the total number of transactions is less than 0, then it will be replaced by 0.

• If the Net Sales <0 then vl = 0. If for some reason (eg. error or returns) the sales value is less, then 0 then it will be replaced by 0.

• If the Total Net Sales <0 then Vl_tot = 0. If for some reason the total net sales value is less than 0 then it will be replaced by 0.

• If the Quantities <0 then qtd = 0. If for any reason the quantity is less than 0, then it will be replaced by 0.

• If the Total Quantities <0, then qtd_tot = 0. If for any reason the total quantity is less than 0, then will be replaced by 0.

• If Discounts <0, then Desconto_em = 0. If for some reason the discount value is less than 0, then it will be replaced by 0.

• If Total Discounts <0, then Desconto_tot = 0. If for some reason the total discount value is less than 0, then it will be replaced by 0.

• If the Number of Products <0, then Num_prod = 0. If the number of products is less than 0, it will then be replaced by 0.

• If the Total Number of Products <0, then Num_prod_tot = 0. If by any reason the total value of products is less than 0, it will then be replaced by 0.

• If the Number of Weeks <0, then Num_semanas = 0. If the number of weeks is less than 0, it will then be replaced by 0.

• If the Total Number of Weeks <0, then Num_semanas_tot = 0. If the total number of weeks is less than 0, it will then be replaced by 0.

Some of the numerical variables with missing or incongruent values cannot be replaced by zero; it is rather preferable that they be replaced by, for instance, the mean, which was the case for, for example, the distance variables. The treatment for this type of situations was as follows:

• If distance to Continent Modelo=. then Distance to Continent Modelo = 6. Corrected by the average.

• If distance to Continent =. then Continent = 53. Corrected by the average.

• If distance to Continente Bom Dia =. then Continente Bom Dia = 68. Corrected by the average.

• IF distance to the Intermarche =. then Intermarche = 45. Corrected by the average.

• If distance to Pingo Doce =. Then Pingo Doce = 10. Corrected by the average.

• If distance to Jumbo =. then Jumbo = 58. Corrected by the average.

• If distance to Lidl =. then Lidl = 44. Corrected by the average.

• If distance to Mini Preço =. then Mini Preço = 44. Corrected by the average.

• If Recency_tot =. then Recency_tot = 365. It will be replaced with the maximum recency value, 365 days.

• If Age <18 then Age = 18. The value of Age is corrected by the minimum age value of an adult.

• If Age> 100 then Age = 100. If the value for Age is above 100, it is corrected for a maximum of 100 years.

• If Age =. then Age = 50. The missing value is corrected by the average.

• If Activation Date =. then Activation Date = 8. The missing value is replaced by the average.

• If Aggregate (family members) =. then Aggregate = 3. Missing value is replaced by the average.

Finally, all the remaining missing values for numerical variables were replaced by 0.

5.2 Data Transformation

The data transformation process used was the standardization of numerical variables, which allows all the variables to be on the same scale. This process grants the comparison of scores between different types of variables, hence data transformation is necessary for all variables to enter the model with the same weight.

To perform the standardization of the variables, the mean and standard deviation were calculated. Then, for each observed value of the variable, the mean was subtracted and divided by the standard deviation.

This process creates standard scores that represent the number of standard deviations above or below the mean to which a specific observation falls.

5.3 Outlier Treatment

As seen in the previous chapter, some variables contain severe outliers, which may bias the estimates of the regression parameters. However, outlier removal from the data must be proceeded with caution since, in many cases, there is a valid reason for the observations to be outliers.

Data transformation was performed before the outlier removal for two reasons: removing outliers before data transformation can cause the data to appear normally distributed, hence conducting data transformation after will not improve the fit and because it is clearer to define where the outlier behavior starts.

Therefore, the standardized values calculated in the previous section were identified as outliers and removed when they assumed greater values than 2 and less than -2, which means that these observations fall at least 2 standard deviations above or below the mean.

5.4 Principal Component Analysis

The Principal Component Analysis (PCA) is a mathematical procedure that considers the total variance in the data and uses an orthogonal transformation to convert a set of observations into a smaller set of linear combinations.

PCA achieves dimension reduction while keeping as much variation as possible by creating new, artificial variables called principal components, which are nothing but a linear combination of the observed variables.

No information about groups is used in the dimension reduction, so PCA brings out the dominant patterns in a dataset. Therefore, PCA is often used to make it easier to explore and visualize large data sets.

Concluding, this analysis was applied to the continuous variables to reduce the number of existent variables without loss of explanatory power. After the application of PCA 14 components were kept which explain about 80% of the variance. Hence, the Logistic Regression will be performed using the 14 Factors and the binary variables mentioned in the previous chapters.

NAME	n_dias	n_dias_ini_mail	per_desconto	RECENCY	IDADE	DATA_ATIVACAO	FAMILY_MEMBERS	var_recency	perc_semanas	var_semana	per_trx
Factor1	-0,177231932	-0,200045356	-0,225485518	-0,470980927	0,007668164	-0,014781352	0,005942624	-0,308558188	0,778408522	0,194632538	0,826873699
Factor2	0,01762624	0,003996936	0,150252445	0,158571124	0,028566208	0,009660146	-0,00019143	0,10442819	-0,214978296	-0,12576229	-0,086982844
Factor3	0,139718576	0,12167557	0,192123302	-0,260726058	0,165675828	0,333981868	0,009772386	-0,002872396	0,420125745	0,151645978	-0,100108544
Factor4	0,07813001	0,058417038	0,043941587	-0,018787775	-0,030215966	0,010779626	0,023434791	0,000172533	0,049977726	0,024700481	0,030487331
Factor5	-0,235680761	-0,301091316	-0,187916629	0,183782915	0,038769956	0,002452687	-0,007314566	0,142931972	-0,14319325	-0,072217499	-0,04277771
Factor6	-0,319218612	-0,3502339	-0,141786471	0,065344916	0,137296287	0,209783418	-0,00398444	0,127168576	0,042914024	0,003792759	-0,141308583
Factor7	0,622631083	0,622380422	0,111782625	0,057951179	-0,02049492	-0,089559089	-0,02587649	0,046547383	-0,062585832	0,022610673	0,041387126
Factor8	-0,189999207	-0,16836774	-0,02272446	0,043141673	-0,119839878	-0,055486172	0,002093671	0,02899022	0,11076092	0,208836386	-0,047593715
Factor9	0,046061615	0,040617687	0,076365968	0,458094421	0,523181828	0,432686446	0,056714234	0,560747198	0,021632322	0,169656181	0,108397463
Factor10	0,215689523	0,143829336	-0,155816322	-0,009605544	-0,029345153	0,289806387	0,118860272	0,102344881	0,17654903	0,028982358	-0,184721675
Factor11	0,025032515	0,014106159	0,014282533	0,334296937	-0,591405226	-0,411536	-0,060413526	0,515370133	0,121479482	0,29959996	-0,034059243
Factor12	-0,050607946	-0,029237754	0,121654909	0,004538107	-0,091620415	-0,040319176	0,976899264	-0,011544792	-0,025627284	0,027427333	0,014679571
Factor13	0,240417693	0,140639932	-0,589633049	-0,023492208	-0,015182256	0,068494298	0,140452405	-0,017919678	0,030576149	-0,206881867	-0,026928878
Factor14	-0,049936785	-0,010530957	0,319955389	0,069589697	-0,131898666	0,01461179	-0,016929687	0,228751321	0,160978594	-0,80975381	0,074800636

NAME	per_vl	per_descontos	per_produtos	per_quant	per_semanas	cesta_med	qm_tot	num_prod_med	NUM_PROD	NUM_TRX
Factor1	0,697117283	0,264741382	0,689604809	0,759886359	0,902153747	0,13332927	0,299794921	-0,039616609	0,730203977	0,759525813
Factor2	0,337821582	0,259260811	0,07492446	0,136573124	-0,140873665	0,90379618	0,459764291	0,290321224	-0,056265141	-0,218010289
Factor3	-0,275424201	0,107837225	-0,49061312	-0,362202944	0,031519742	0,065036687	-0,113657362	-0,5397698	0,068384073	0,414069144
Factor4	-0,006699268	0,020417219	-0,052320432	-0,019253642	0,037981184	0,021592242	-0,00307997	-0,087089546	-0,00830825	0,051226033
Factor5	-0,119752412	0,345452472	0,151519501	0,078846615	-0,083732835	-0,266762905	0,19180186	0,298329225	0,048241468	-0,150450397
Factor6	-0,001997492	-0,143814235	0,049129085	-0,005472138	-0,094672308	0,083997109	0,050846457	0,175507213	0,165429141	0,049754995
Factor7	0,14643762	0,207620504	0,121063756	0,22277951	0,015820794	-0,119774319	0,118054009	-0,04529037	-0,103521315	-0,055751815
Factor8	-0,058665814	0,407409545	-0,145122351	-0,112091389	0,000455264	-0,040604152	-0,150553386	-0,237670008	0,001474473	0,139742435
Factor9	0,150810587	0,036284442	0,105472344	0,114888287	0,062183349	-0,054560647	-0,151436782	-0,087940749	-0,031178873	0,047271659
Factor10	-0,279085111	-0,217290281	-0,119981348	-0,098859303	-0,087609821	0,08849053	0,502531999	0,30025152	0,331386881	0,190729205
Factor11	-0,002442059	-0,082825761	-0,004767561	-0,041299455	0,010508305	0,024219788	-0,069239677	0,051255699	0,177068305	0,170359597
Factor12	0,029582658	0,003599055	0,00575394	0,036910117	0,005681168	-0,020476349	0,006766761	-0,049509613	-0,05273991	-0,023948269
Factor13	-0,002486121	0,182251474	0,098562784	-0,188519516	-0,003571258	0,039996138	-0,412557489	0,257273384	0,209430849	0,012530269
Factor14	-0,045122326	0,068597752	-0,040398939	-0,031957532	0,135871694	-0,055113522	-0,045816274	0,006290425	0,074421483	0,068699378

NAME	preco_med_tot	desconto_vend	desconto_med	per_toto	taxa_reb_mail_cli	expetativa_desconto	dist_continente	dist_concorr	var_dist	index_conco
Factor1	-0,118501553	0,027082901	0,045703004	-0,089932513	0,081924463	0,067967763	0,021990163	-0,047215253	-0,043235207	-0,015492819
Factor2	0,624164176	0,321663234	0,550798948	-0,093668966	0,036808371	0,904694072	0,027686531	-0,023593584	-0,025467099	-0,00254537
Factor3	0,156261883	0,344966969	0,256359966	-0,328165282	0,665588286	0,115961646	-0,276066398	-0,028875666	-0,011700488	0,226705781
Factor4	0,028033244	0,026846347	0,031008785	-0,003438329	0,035109385	0,032043313	0,290753393	0,908804894	0,890882564	-0,405992209
Factor5	-0,45427623	0,595288322	0,399224479	-0,373600005	0,27577663	-0,301397775	0,011533927	0,053120581	0,051403916	0,024385294
Factor6	0,048807908	-0,184221312	-0,122662969	0,188750126	-0,078333504	0,042845839	-0,661383671	0,265772303	0,314140563	0,594511258
Factor7	-0,23212792	0,156178052	0,122992688	0,121222347	-0,162765131	-0,088805619	-0,380399404	0,092912694	0,118976868	0,372329271
Factor8	0,051345877	0,308587159	0,309728739	0,657187993	-0,467339969	-0,035304137	0,008847761	-0,036379686	-0,034181938	-0,073081313
Factor9	0,069957173	-0,019832377	-0,10131569	0,021863502	-0,018426602	-0,038423173	0,130811081	-0,044680212	-0,057984014	-0,133812495
Factor10	-0,300032065	-0,085943792	0,090482971	0,274138549	-0,04357376	0,045480766	0,136723288	-0,060467079	-0,070567352	-0,18268287
Factor11	0,082506313	-0,045831105	-0,058778427	-0,131576046	0,145479354	0,025439451	-0,035175025	-0,005642256	-0,000689714	0,036034
Factor12	-0,021095503	-4,21532E-06	-0,008330669	-0,035502283	0,003205214	0,007771854	-0,036074695	-0,001150843	0,001411779	0,039814292
Factor13	0,33274265	0,088140153	-0,039283163	-0,010848268	0,029145782	-0,093367295	0,00962506	0,002029986	0,002013589	-0,006647057
Factor14	-0,034809807	0,017593766	-0,034271122	0,113859853	-0,003809494	0,020842356	-0,009602532	-0,003961908	-0,004205372	0,004136185

Table 5.1- PCA

6 Modelling

6.1 General Test Design

In this chapter, the first task was to partition of the modeling base in three parts: 60% for training base and 15% for the validation base and 25% test set.

The training set is used to fit the models; the validation set is used to check the performance of the model so there is the possibility to optimize the system and validate the model (leaving no doubt of overfitting); the test set is used for assessment of the generalization error of the final chosen model.

The split into train/validation/test was based on time and performed by this order (the test set should be the most recent part of data) to avoid the look-ahead bias. The Look-ahead bias occurs when data that would not be known or would not be available during the period in analysis is being used, which leads to inaccurate results. In fact, time-series data requires a simulation where, after training the model, the data coming after the time of creation of the model is evaluated. Therefore, practicing random sampling to split the data does not guaranty accurate results.

The reasoning behind this data split has also to do with the way mailing campaigns are conducted in Sonae: there are either monthly or bimonthly campaigns. As such, the training set embodies two bimonthly mailing campaigns, the validation set a monthly one and, finally, the test set is composed of a bimonthly campaign.

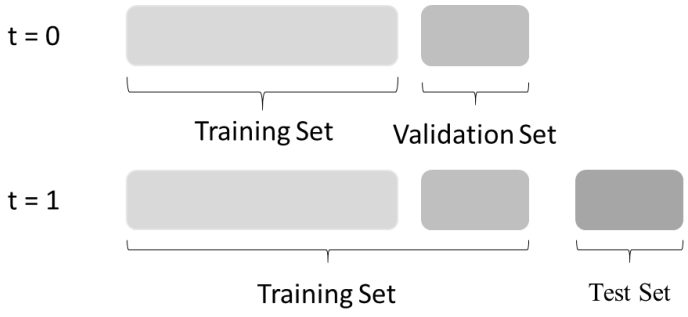

Figure 6.1 - Scheme on the partitioning of the data.

After the construction of this final database, the data was divided into 3 parts: training data with a total of 8307851 rows, the validation database with a total of 1886668 rows and the test set with a total of 3940890 rows.

6.2 Modelling Technique Selection

6.2.1 Logistic Regression

In this step of the project, several Logistic Regression assumptions were considered in order to implement it, including the data preparation processes discussed in the previous chapter. Through SAS code, using PROC LOGISTIC, the Logistic Regression was executed. Of all the existing methods of selection of variables available (None, Backward, Forward and Stepwise), the selected one was **None (Main Effects)** which, by default, implies that all input variables are candidates to be included in the final model.

In table 6.1, the resulting coefficient estimates of the Logistic Regression can be seen, as well as the Wald test.

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.1817	0.0226	34195.9971	<.0001
Factor1		1	0.0605	0.00468	167.1191	<.0001
Factor2		1	0.4995	0.00374	17879.6822	<.0001
Factor3		1	0.00684	0.00360	3.5986	0.0578
Factor4		1	0.0129	0.00417	9.5794	0.0020
Factor5		1	0.2323	0.00202	13245.6493	<.0001
Factor6		1	-1.8307	0.00457	160595.966	<.0001
Factor7		1	-0.0291	0.00344	71.3840	<.0001
Factor8		1	0.0699	0.00328	454.7668	<.0001
Factor9		1	-0.1237	0.00347	1268.6763	<.0001
Factor10		1	0.0509	0.00271	351.1424	<.0001
Factor11		1	0.0296	0.00298	98.2019	<.0001
Factor12		1	0.5429	0.00437	15427.9910	<.0001
Factor13		1	-0.0877	0.00332	696.6675	<.0001
Factor14		1	-0.00190	0.00297	0.4097	0.5221
TIPO_CUPAO	A	1	-1.8476	0.0148	15550.4437	<.0001
TIPO_CUPAO	C	1	1.8015	0.00935	37105.4249	<.0001
LIMITE		1	-2.4515	0.0295	6888.1004	<.0001
BEBE		1	-2.5266	0.3341	57.1808	<.0001
MERCEARIA SALGADA		1	-0.6858	0.00994	4758.8033	<.0001
MERCEARIA DOCE		1	-0.9242	0.0111	6876.7635	<.0001
LACTICINIOS		1	-0.9267	0.0145	4108.9959	<.0001
CONGELADOS		1	-1.7175	0.0269	4074.0155	<.0001
LIMPEZA DO LAR		1	0.0232	0.0120	3.7400	0.0531
HIGIENE E BELEZA		1	-2.3466	0.0205	13066.3126	<.0001
REGION	AÇORES	1	1.1190	0.0952	138.1097	<.0001
REGION	CENTRO	1	-0.2181	0.0201	118.1706	<.0001
REGION	MADEIRA	1	-0.2036	0.0240	72.2107	<.0001
REGION	NA	1	-0.2646	0.0212	155.8266	<.0001
REGION	NORTE	1	-0.2006	0.0202	98.4637	<.0001
flag_comp_p1		1	-0.3531	0.0137	667.0255	<.0001
segmento_val	frequente	1	0.2421	0.0100	583.8807	<.0001
segmento_val	leal	1	0.2531	0.0109	536.5774	<.0001
segmento_val	no_value	1	-0.5193	0.0275	357.2158	<.0001

Tabela 6.1- Coefficient estimates and Wald test.

Wald test is a hypothesis test used to access whether the independent variables are significantly associated with the response variable or not. All the effects with a p-value > 0.05 (95% significance) were rejected, those being Factor 3, Factor 14 and the flag Limpeza do Lar.

6.2.2 Stochastic Gradient Descent

As mentioned in the State of the Art chapter, Stochastic Gradient Descent refers to calculating the derivative from each training data instance and calculating the update immediately. As same as when applying Logistic Regression, while modelling SGD several assumptions also must be considered, including the data preparation processes discussed in the previous chapter.

One of the most important steps to consider while implementing SGD is the choice of the learning rate η since this is the parameter which determines how fast or slow the moves are made towards the optimal weights. Hence, if η is too large the optimal solution will be skipped, otherwise if it is too small, it will require too many iterations to converge to the best values. Thus, using a good learning rate is crucial.

Furthermore, the learning rate value must be a small real value. Different values were experimented before reaching a final decision, as seen in 6.1. By analyzing the graph, it is noticeable that the learning rate which provides the minimum average error is in between 0,1 and 0,001, hence the chosen rate was 0,01.

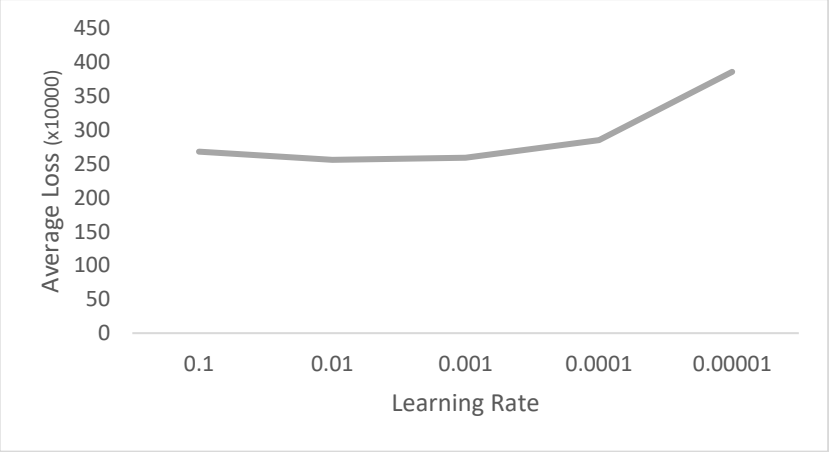

Figure 6.2- Average cost plot for different learning rate values.

Another important decision is to decide how many passes, epochs or loops through the training dataset are needed. Literature suggests that SGD often does not require more than

1-to-10 epochs to converge on good or good enough coefficients. The model was assessed for 1, 10 and 15 loops.

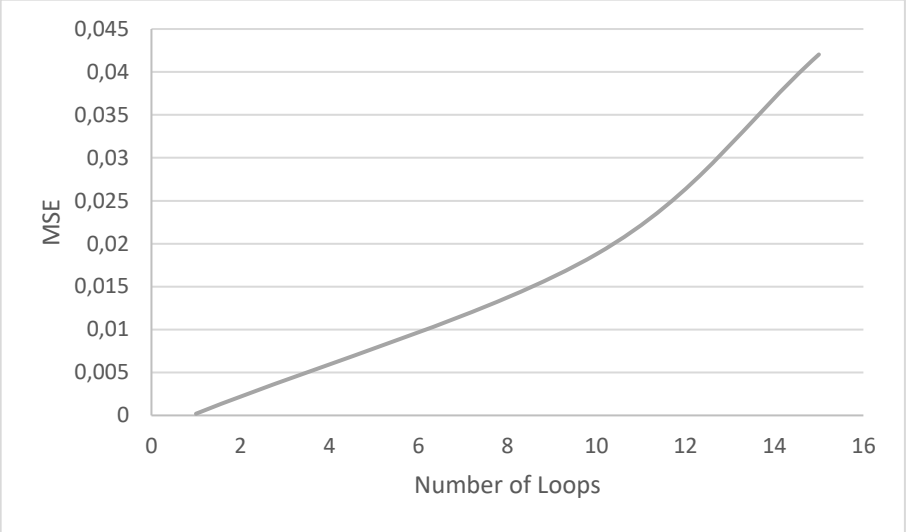

Figure 6.3- Mean Squared Error by Number of Epochs

By analyzing the Figure 6.2 it is noticeable that at a learning rate of 0,01, the mean squared error increases with higher number of epochs. This happens because the final estimated coefficients for a SGD with more than one epoch and a constant learning rate are calculated as the average of coefficient estimates in the last epoch. The SGD is more efficient if the algorithm runs only once over the data (Zhang, 2004) hence the chosen number of epochs was 1.

6.3 Comparison of Results

After parametrizing and adjusting the models, the results obtained in the validation process were evaluated for Continente's existent model, the Logistic Regression before the SGD implementation and, finally, the Stochastic Gradient Descent. This comparison is necessary to infer if the chosen model meets both the technical and business goals, so it is later used for the scoring and operationalization in the business.

It must be noted that Continente's existent model is also a Logistic Regression with the parameters initially calculated when the model was created for Sonae, and the Logistic Model refers to the logistic regression created in the scope of this dissertation. Thus, poorer results from the existing model must be analyzed carefully, since the reason for that might have to do with the model's use of older data and, therefore, for being less adapted to the current context.

To evaluate these models both the technical aspects and some more subjective aspects must be considered.

As mentioned in chapter 2, the pre-choice of the methodology to apply in this dissertation regarded the interpretability and operationalization of the final model in the business: Logistic Regression and SGD provide specific information on each parameter of the forecast and allow the analysis of the individual coefficients of the logistic equation. In addition, the fact that an equation is obtained facilitates the implementation and operationalization of these models in other business teams. Finally, their simplicity of use and easily interpretable results are of extreme importance in retail business, since the perception of which variables are contributing to a particular behavior and how they are linked or related influence the decision making-process.

Considering the abovementioned nuances, the statistical comparison of the models was generated in the validation data set.

Since the target variable of this problem is unbalanced, with two groups of 96% and 4%, to measure and compare models, guiding the evaluation of the models only by Accuracy could be misleading since the target value does not have an equal number of samples belonging to each class. As an alternative, the Lift metric computes the ratio between target response and average response, thus it allows the perception of how much better the model is when comparing to a random forecast. Thus, success is achieved when the lift is > 1, which means that the probabilities of the occurrence of the input value and of the variable to predict are dependent on one another and makes those rules potentially useful for predicting the later in future data sets. As such, the main criteria of evaluation will be the Lift Value.

Although the final goal of the project is not to obtain a clear customer separation (those who redeem and those who do not), but instead to predict the Coupon Redemption Rate; the Accuracy, Specificity and Sensibility metrics were extracted from the confusion matrix of each model and consulted.

MODEL	DATA	FN	TN	FP	TP
Existent Model	Validation set	52943	788729	3125	8457
Logistic Regression	Validation set	32606	248286	7414	12782
SGD	Validation set	29960	780008	11846	31440

Tabela 6.2- Confusion Matrix

The Confusion table is a matrix which describes the complete performance of the model. Assuming a binary problem in which there are two possible predicted classes: "YES" and "NO" by determining the True Positives (TP): The cases in which YES was predicted and the actual output was also YES, the True Negatives (TN): The cases in which NO was predicted and the actual output was YES, the False Positives (FP): The cases in which YES was predicted and the actual output was NO and the False Negatives (FN): The cases in which NO was predicted and the actual output was also NO (Gama, 2012).

All the metrics mentioned above result of the combination of the confusion table's outputs, as it can be stated on table 6.3.

Acuracy	Missclassification Rate	Sensitivity	Specificity
$\frac{TP+FN}{TN+TP+FN+FP}$	1- Accuracy	$\frac{TP}{TP+FN}$	$\frac{TN}{TN+FP}$

Table 6.3- Metric formulas

MODEL	LIFT	ACCURACY	MISCLASSIFICATION RATE	SENSITIVITY	SPECIFICITY
Existent Model	4,19842	0,867082	0,065710797	0,137736156	0,003946435
Logistic Regression	10,14712	0,934289	0,132917951	0,281616286	0,028994916
SGD	10,09358	0,951004	0,048995961	0,512052117	0,014959828

Tabela 6.4- Model Results

It is noticeable by analyzing table 6.4 that the Lift value in Logistic Regression and SGD is far superior to the existent model, which was the main criterion of model selection. These models are around 10 times better than random forecasting.

In addition, SGD has the highest Accuracy value, thus also the lowest value of Misclassification Rate, which measures the proportion of poorly classified instances. It also registers the maximum value of Sensitivity which corresponds to the proportion of positive data points which are correctly considered positive, with respect to all positive data points. Logistic Regression, on the other side has the highest Specificity which corresponds to the proportion of negative data points that are mistakenly considered positive, with respect to all negative data points.

7 Evaluation/Validation

At this stage, the model was scored using the test set. To report on the model's performance on the independent test set, we analyze two campaign months: January and February 2018, consisting of 113 promotional coupons.

To obtain the results, the evaluation of the final model was computed using two methods: in method A the sum of the predicted redemptions decided by threshold of redemption or no redemption was used, while in method B the sum of the predicted probabilities attributed to each client by the model is used. These letters were attributed to facilitate the interpretation of the following tables of results.

The metric used to evaluate the models was the Mean Squared Error (MSE), which computes the average of the square of the difference between the original values and the predicted values. The advantage of MSE is that by computing the square of the error the effect of larger errors become more pronounced then smaller errors. The MSE is given by the following formula:

$$MeanSquaredError = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$$

The table 7.1 depicts the MSE results obtained for the redemption rate prediction of promotional coupons for all the models under analysis in both methods.

MODELO	MSE (A)	MSE (B)
Existent Model	0,73%	0,22%
Logitic Regression	0,67%	0,28%
SGD	0,32%	0,02%

Table 7.1- Mean Squared Error

As expected, the MSE values of method B (the sum of the predicted probabilities attributed to each client) surpass the results of method A (sum of the predicted

redemptions decided by threshold) in every model. SGD registers the smallest MSE values.

MODELO	MAX ε (A)	MAX ε (B)	MIN ε (A)	MIN ε (B)
Existent Model	13,80%	4%	0%	0,23%
Logistic Regression	12,70%	6,10%	0%	0%
SGD	8,50%	10%	0,10%	0%

Table 7.2- Maximum and Minimum error (ε) for each model

Observing table 7.2, it is perceivable that SGD has the lowest maximum error value ε for method A, but also the highest for method B and the highest minimum error value for method A.

MODEL	LIFT	ACCURACY	MISCLASSIFICATION RATE	SENSITIVITY	SPECIFICITY
Existent Model	0,770038	0,499945	0,500055074	0,000368698	0,000588887
Logistic Regression	10,87527	0,934289	0,045008874	0,035264484	0,001736012
SGD	12,17004	0,9555952	0,044047559	0,115287067	0,004494507

Tabela 7.3- Final Model result

It is noticeable by analyzing table 7.3 that SGD surpasses the other models in every metric: it registers the highest Lift value, highest Accuracy and Sensitivity. On the other hand, it registers the lowest values of Misclassification Rate and Specificity.

In turn, the Existent model has the poorest scores, with an emphasis to the misclassification rate, thus Accuracy also of around 50%.

Thusly, assessing all these metrics, it is indisputable that SGD introduced improvements to the predictive model.

8 Deployment

Nowadays, scoring the existent model for the partner Continente is a 4-day process, in which each coupon takes an average of 10 minutes to be scored. The additional time needed to deploy SGD in the business model has to be weighed since the solution to any business problem will require it to be closed-loop with the time available between event and action.

In fact, in the development of this project, we only considered a sample of 10% of the clients and performed the model for coupons of the "Alimentar" category, disregarding "Frescos" and "Não Alimentar" coupons. The fact that Continente is only one of the partners of Continente Card must also be considered since the overall process is even more time-consuming.

The following table represents the computational time in minutes of the models performed on the scope of this dissertation vs. the MSE results obtained, illustrating the fact that even though SGD implementation is more time-costly, there are performance gains comparing to the other less time-consuming models.

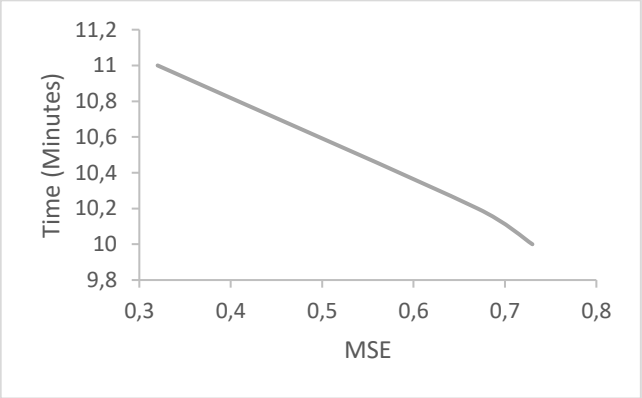

Figure 8.1 - Time vs. MSE

Likewise, the Continent Card's Management Board must compare the computation time and memory costs associated with SGD with the gains in model performance and decide whether the application in the business context and the scalability of these processes for the other Card Partners is deployable or not.

9 Conclusions

9.1 Final Considerations

This dissertation's outline was the creation of an Adaptive Learning Redemption Rate Prediction Model for Continente, the largest Partner of Continente Card.

The Success criteria defined for Data Mining and Business goals were successfully achieved, since the prediction of the redemption rate of promotional mailing coupons in an Adaptive Machine Learning framework was successfully implemented and outperformed the results of the Existent Model and the Logistic Regression Model. The business can now apply the adaptive model to predict the redemptions rates for future mailing coupons and perform the scalability of the model for other Partners of the Continente Card, and thus optimize the management of stocks, the sales forecast and carry out strategic adjustments if necessary.

On a personal level, the development of this dissertation allowed consolidation and knowledge expansion of data analysis, modelling and extraction of knowledge from data. Besides the additional literary and theoretical knowledge, it allowed the development of programming experience in SAS language.

On a professional level, the development of this dissertation in Sonae MC granted the opportunity to explore a real size database and obtain broader knowledge of Sonae's business structure and to explore the configurations of the promotional, demographic and transactional Continente Card's data. In addition, it also granted the possibility of studying and getting involved with several of Sonae MC's business areas and stages of conducting direct marketing campaigns.

9.2 Limitations and Future Steps

Even though the Data Mining and Business goals were successfully achieved, some limitations and future grounds of work must be considered.

One of the major limitations of this model regards the implementation of PCA, which reduced the interpretability of the parameters in the modelling phase.

Regarding possible model improvements, the distance to competitor's variable was included even though it is manifestly incomplete, and its information does not correspond to reality. Hence, it could be improved by implementing, for instance, the Van der Waals forces to consequently further improve the model.

In addition, it might be relevant to test the inclusion of a channel of contact variable since, as mentioned before, the customers receive the coupons by different means of contact- such as letter, newsletter and mobile app- and might respond differently to each stimulus.

10 References

Anderson, T. W., Anderson, T. W., Anderson, T. W., Anderson, T. W., & Mathématicien, E. U. (1958). *An introduction to multivariate statistical analysis* (Vol. 2, pp. 5-3). New York: Wiley.

Baesens, B., Viaene, S., Van Den Poel, D., Vanthienen, J., & Dedene, G. (2002). Bayesian neural network learning for repeat purchase modelling in direct marketing. *European Journal of Operational Research*, 138(1), 191–211.

Bateson, J. E., & Hoffman, K. D. (1997). Essentials of services marketing.

Basheer, I. a., & Hajmeer, M. (2000). Artificial neural networks: Fundamentals, computing, design, and application. *Journal of Microbiological Methods*, *43*(1), 3–31. doi:10.1016/S0167-7012(00)00201-3

Bryant, C. H., & Muggleton, S. H. (2000). Closed loop machine learning. *REPORT-UNIVERSITY OF YORK DEPARTMENT OF COMPUTER SCIENCE YCS*.

Bottou, L. (2012). Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade* (pp. 421-436). Springer, Berlin, Heidelberg.

Coelho, V. N., Oliveira, T. A., Coelho, I. M., Coelho, B. N., Fleming, P. J., Guimarães, F. G., Ramalhinho, H., Souza, M. J., Talbi, E.-G., & Lust, T. (2017). Generic Pareto local search metaheuristic for optimization of targeted offers in a bi-objective direct marketing campaign. *Computers & Operations Research*, 78, 578–587.

Cui, G., Wong, M. L., & Lui, H.-K. (2006). Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming. *Management Science*, 597–612.

Cullinan & GJ (1977). Picking them by their batting averages' recency-frequency-monetary method of controlling circulation. *Direct Mail/Marketing Association. N.Y.*, Manual rel.

Davenport, T. H., Harris, J. G., & Morison, R. (2010). *Analytics at work: Smarter decisions, better results*. Harvard Business Press.

Deichmann, J., Eshghi, A., Haughton, D., Sayek, S., & Teebagy, N. (2002). Application of multiple adaptive regression splines (mars) in direct response modeling. *Journal of Interactive Marketing*, 16(4), 15–27.

Du, W., & Zhan, Z. (2002). Building decision tree classifier on private data. *Proceedings of the IEEE International Conference on Privacy, Security and Data Mining-Volume 14*, 1–8. Retrieved from http://portal.acm.org/citation.cfm?id=850784

Eisenstein, E. M., & Lodish, L. (2002). Marketing decision support and intelligent systems: precisely worthwhile or vaguely worthless?. *Handbook of marketing*, 436-455.

Elsner, R., Krafft, M., & Huchzermeier, A. (2004). The 2003 ISMS Practice Prize Winner: Optimizing Rhenania's Direct Marketing Business Through Dynamic Multilevel Modeling (DMLM) in a Multicatalog-Brand Environment. *Marketing Science*, 23(2), 192–206.

Finlay, S. (2014). *Predictive Analytics, Data Mining, and Big Data. Miths, misconceptions and methods*. Chennai, India: Palgrave Macmillan.

GAMA, João; CARVALHO, André Ponce de Leon; FACELI, Katti; LORENA, Ana Carolina; OLIVEIRA, Marcia. (2015) *Extração de conhecimento de dados: data mining.* 2. ed. Lisboa: Edições Sílabo, 2015. 428 p.

Gepperth, A., & Hammer, B. (2016). Incremental learning algorithms and applications. In *European Symposium on Artificial Neural Networks (ESANN)*.

Granja, P. (2017) *Previsão da Taxa de Rebate de Cupões Promocionais em Marketing Direto*. Dissertação (Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão) – FEP, Universidade do Porto.

Hamilton, H. J., Gurak, E., Findlater, L., Olive, W., & Ranson, J. (2012). Computer Science 831: Knowledge Discovery in Databases. Retrieved from http://www2.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/4_dtrees1.html.

Hammer, B., & Toussaint, M. (2015). Special issue on autonomous learning.

Haykin, S. (1999). Neural Networks-A Comprehensive Foundation.

Hosmer, D. W. & Lemesbow, S. (2000). Applied logistic regression (2nd ed.). John Wiley & Sons.

Hsu, M. J., & Ho, C. P. (2012). Creating a knowledge discovery model using MOEX's examination database for in-depth analysis and reporting. In *Proceedings - 2012 IEEE Symposium on Robotics and Applications, ISRA 2012* (pp. 705–707). Kuala Lumpur. doi:10.1109/ISRA.2012.6219288

Javaheri, S. H., Sepehri, M. M., & Teimourpour, B. (2013). *Response Modeling in Direct Marketing. A Data Mining-Based Approach for Target Selection.* Elsevier Inc.

Jiawei, H., Kamber, M., Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*.

Kim, Y. & Street, W. N. (2004). An intelligent system for customer targeting: A data mining approach. *Decision Support Systems*, 37(2), 215–228.

Kotler, P. (1989). From mass marketing to mass customization. *Planning review*, *17*(5), 10-47.

Kotler, P., & ARMSTRONG, G. (2003). Princípios de marketing. Tradução de: Arlete Simille Marques e Sabrina Cairo.

Kumar, P. R., & Varaiya, P. (2015). *Stochastic systems: Estimation, identification, and adaptive control* (Vol. 75). SIAM.

Lemos, E. P., Steiner, M. T. A., & Nievola, J. C. (2005). *Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining*. São Paulo.

Ling, C. & Li, C. (1998). Data Mining for Direct Marketing: Problems and Solutions. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 98, 73–79.

Maimon, O., & Rokack, L. (2005). *The Data Mining and Knowledge Discovery Handbook.* Israel: Springer.

Mathews, V. J., & Xie, Z. (1993). A stochastic gradient adaptive filter with gradient adaptive step size. *IEEE Transactions on Signal Processing*, *41*(6), 2075-2087.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, 22–31.

Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology. In *Proceedings of European Simulation and Modelling Conference-ESM'2011* (pp. 117-121). Eurosis.

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of Statistical Analysis and Data Mining Applications*. California: Elsevier Inc.

Olson, D. L., Cao, Q., Gu, C., & Lee, D. (2009). Comparison of customer response models. *Service Business*, *3*(2), 117-130.

Olson, D. L. & Chae, B. (2012). Direct marketing decision support through predictive customer response modeling. *Decision Support Systems*, 54(1), 443–451.

Poel, D. V. D. (2003). Predicting Mail-Order Repeat Buying : Which Variables Matter ? *Tijdschrift voor Economie en Management*, XLVIII, 371–404.

Potharst, R., Kaymak, U., & Pijls, W. (2001). Neural Networks for Target Selection in Direct Marketing. In *Neural Networks in Business* (pp. 89–111). IGI Global.

Ramaswamy, V., Srinivasan, S. S., & Srini, S. (1998). Coupon Characteristics and Redemption Intentions: A Segment-Level Analysis. *Psychology & Marketing*, 15(January 1998), 59–80.

Rico-Juan, J. R., & Iñesta, J. M. (2014). Adaptive training set reduction for nearest neighbor classification. *Neurocomputing*, *138*, 316-324.

Ruvolo, P., & Eaton, E. (2013, February). ELLA: An efficient lifelong learning algorithm. In *International Conference on Machine Learning* (pp. 507-515).

Sen, A. & Srivastava, M. (1990). *Regression Analysis: theory, methods, and applications*.

Setnes, M. & Kaymak, U. (2001). Fuzzy modeling of client preference from large data sets: an application to target selection in direct marketing. *IEEE Transactions on Fuzzy Systems*, 9(1), 153–163.

Shalizi, C. R. (2012). *Advanced Data Analysis from an Elementary Point of View*. Pittsburgh.

Shin, H. & Cho, S. (2006). Response modeling with support vector machines. *Expert Systems with Applications*, 30(4), 746–760.

Suh, E., Noh, K., & Suh, C. (1999). Customer list segmentation using the combined response model. *Expert Systems with Applications*, 17(2), 89–97.

Vaughan, I. P., & Ormerod, S. J. (2003). Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. *Conservation biology*, *17*(6), 1601-1611.

van Geloven, S. (2002). Combining Target Selection Algorithms in Direct Marketing.

Verhoef, P. C., Spring, P. N., Hoekstra, J. C., & Leeflang, P. S. H. (2002). The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands. *Decision Support Systems*, 34(4), 471–481.

Viaene, S., Baesens, B., Poel, D. V. D., Dedene, G., & Vanthienen, J. (2001). Wrapped input selection using multilayer perceptrons for repeat-purchase modeling in direct marketing. *International Journal of Intelligent Systems in Accounting, Finance & Management*, 10(2), 115–126.

Ville, B. (2006). Decision Trees - What Are They? In *Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner*. SAS Institute Inc.

Wong, M. L. (2016). Evolutionary Programming Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming Hon-Kwong Lui. 52(4), 597–612.

Wooldridge, J. M. (2013). *Introductory Econometrics. A modern approach.* (S.-W. C. Learning, Ed.). Michigan.

Yu, E. & Cho, S. (2006). Constructing response model using ensemble based on feature subset selection. *Expert Systems with Applications*, 30(2), 352–360.

Zhang, T. (2004, July). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning* (p. 116). ACM.